Jurnal Psikologi
Volume 46, Nomor 2, 2019: 163 – 180
DOI: 10.22146/jpsi.24260

ISSN 0215-8884 (print)
ISSN 2460-867X (Online)
https://jurnal.ugm.ac.id/jpsi

Pemahaman Peneliti dan Mahasiswa Psikologi mengenai Besaran Sampel: Data dan Simulasi

On How Indonesian Psychological Researchers and Students Understand Sample Size Calculation: Data and Simulation

Wisnu Wiradhany[1], Krisna Adiasto[2], Jony Eko Yulianto[3,4], & Indra Y. Kiling[5]

[1]Departemen Psikologi, Fakultas Ilmu Perilaku dan Ilmu Sosial, University of Groningen, [2]Behavioural Science Institute, Radboud University, [3]Fakultas Psikologi, Universitas Ciputra, [4]School of Psychology, Massey University, [5]Prodi Psikologi, Fakultas Kesehatan Masyarakat, Universitas Nusa Cendana

Abstract. The lack of knowledge on how to determine sample sizes in experiments is arguably one of the main reasons underlying the replication crisis in psychological science. We distributed a survey to Indonesian students and researchers concerning 1) familiarity and understanding of statistical concepts related to sample size determination, 2) current sample size determination practices in experiments, and 3) ideal sample sizes for experiments. Subsequently, we simulated expected statistical power given the sample sizes reported in the survey. The results demonstrated that 1) while a majority of participants were somewhat familiar with statistical concepts related to sample size determination, they did not always endorse the correct and/or complete definition of each concept. Furthermore, 2) our participants relied on practical considerations in determining sample sizes. Consequently, 3) the reported sample sizes did not have sufficient power to detect small to medium effect sizes, which are commonly present in psychological science.

Keywords: effect size; replication crisis; sample size; statistical power

Abstrak. Salah satu kendala utama yang melatarbelakangi krisis replikasi dalam psikologi adalah kurangnya pemahaman peneliti psikologi mengenai besaran sampel ideal dalam penelitian. Survei disebarkan kepada 354 mahasiswa dan peneliti psikologi untuk menanyakan 1) familiaritas dan pemahaman mengenai konsep-konsep statistika terkait penentuan besaran sampel, 2) kebiasaan menentukan besaran sampel penelitian, dan 3) pendapat mengenai besaran sampel ideal. Selanjutnya, menggunakan besaran sampel yang dilaporkan untuk mensimulasikan capaian kekuatan uji statistik (*power*). Hasil penelitian menunjukkan bahwa 1) sebagian besar mahasiswa dan peneliti psikologi familiar dengan konsep-konsep statistika yang erat kaitannya dengan penentuan besaran sampel, namun memiliki pemahaman yang keliru dan/atau tidak utuh mengenai konsep-konsep tersebut, dan 2) mereka mengandalkan pertimbangan pragmatis dalam menentukan besaran sampel. Akibatnya, 3) besaran sampel yang dianggap ideal tidak memiliki kekuatan uji statistik yang cukup untuk mendeteksi besaran efek kecil dan sedang, yang umum ditemukan dalam penelitian-penelitian psikologi.

***Kata kunci*:** besaran efek; besaran sampel; kekuatan uji statistik; krisis replikasi

[1] Korespondensi mengenai artikel ini dapat melalui: w.wiradhany@rug.nl

Hasil penelitian kolaboratif *Open Science Collaboration* yang diterbitkan di jurnal ilmiah *Science* melaporkan bahwa penelitian Psikologi dalam jurnal ilmiah berindeks tinggi yang hasilnya dapat direplikasi hanya mencapai 36%. Artinya, hampir dua per tiga dari temuan ilmiah yang dilaporkan tidak dapat direproduksi secara independen (Open Science Collaboration, 2015). Artikel tersebut membuktikan bahwa Psikologi sebagai bidang ilmu juga mengalami krisis replikasi dalam sains (Button *et al*., 2013; Ioannidis, 2005; Lindsay, 2015, untuk pembahasan krisis replikasi dalam sains), dan menambah panjang daftar bidang keilmuan yang hasil penelitiannya tidak dapat direplikasi (Button *et al*., 2013, untuk replikasi penelitian dalam neurosains, dan Ioannidis, Ntzani, Trikalinos, & Contopoulos-Ioannidis (2001) untuk replikasi penelitian dalam biologi molekuler). Temuan ilmiah yang tidak dapat direplikasi memiliki dampak yang fatal. Bukan saja sumber daya finansial yang besarannya tidak sedikit terbuang sia-sia, hasil penelitian yang tidak dapat direplikasi juga berpotensi menimbulkan masalah ketika digunakan sebagai acuan pembuatan kebijakan. Sebagai contoh, dalam bidang kedokteran dan farmasi, hasil penelitian yang tidak dapat direplikasi berpotensi menghasilkan obat-obatan yang memiliki dampak negatif (Simonsohn, Nelson, & Simmons, 2014).

Secara umum, krisis replikasi dalam sains dan, secara spesifik, dalam psikologi dilatarbelakangi dua jenis masalah. Jenis masalah pertama meliputi praktik-praktik yang melanggar kaidah etika penelitian (*Questionable Research Practices;* QRP). Praktik-praktik tersebut mencakup bias publikasi (*publication bias*), yakni anggapan bahwa penelitian yang melaporkan perbedaan signifikan secara statistik memiliki kemungkinan lebih besar untuk dipublikasikan di jurnal-jurnal ilmiah (Ioannidis, 2005; Ioannidis, Munafò, Fusar-Poli, Nosek, & David, 2014); *cherry picking,* yakni dengan sengaja hanya melaporkan hasil eksperimen yang signifikan secara statistik; dan *p-hacking,* yakni melakukan analisis inferensial berulang-ulang setiap pengumpulan data hingga menemukan hasil yang signifikan secara statistik (Rouder, 2014; Simmons, Nelson, & Simonsohn, 2011). Praktik-praktik di atas menimbulkan "bias positif" terhadap temuan-temuan yang dilaporkan dalam artikel ilmiah. Artinya, artikel-artikel ilmiah yang dipublikasikan cenderung melaporkan temuan positif (signifikan secara statistik) secara berlebihan dan/atau tidak proporsional secara matematis (Ioannidis, 2005). Perlu diketahui bahwa rasio temuan positif dan temuan negatif yang dilaporkan terus meningkat dalam setengah abad terakhir (Ioannidis *et al*., 2014).

Jenis masalah kedua bersifat lebih teknis, yakni menyangkut minimnya pengetahuan peneliti mengenai analisis statistik inferensial (Badenes-Ribera, Frias-Navarro, Iotti, Bonilla-Campos, & Longobardi, 2016; Badenes-Ribera, Frías-Navarro, Monterde-I-Bort, & Pascual-Soler, 2015; Gigerenzer, 2004; Hoekstra, Morey, Rouder, & Wagenmakers, 2014) dan eksploitasi statistik inferensial untuk mendapatkan temuan positif (Simonsohn *et al*., 2014). Secara spesifik, kurangnya pemahaman peneliti mengenai uji statistik inferensial menyebabkan sebagian besar temuan ilmiah memiliki kekuatan uji statistik (*statistical power*) yang rendah, sebagai konsekuensi penentuan besaran sampel penelitian yang tidak ideal (Bakker, Hartgerink, Wicherts, & van der Maas, 2016; Button *et al*., 2013). Padahal, kekuatan uji statistik, yakni probabilitas jangka

panjang untuk menyimpulkan bahwa suatu efek tidak ada ketika efek tersebut benar tidak ada untuk besaran efek (*effect size*) dan jumlah sampel tertentu (Cohen, 1992a), berkaitan erat dengan probabilitas jangka panjang mereplikasi temuan ilmiah (Button *et al*., 2013). Kekuatan uji statistik memiliki relasi yang erat dengan besaran sampel dan besaran efek. Deteksi besaran efek kecil membutuhkan sampel besar untuk mencapai kekuatan uji yang tinggi, dan demikian pula sebaliknya (lihat Gambar 1).

Artikel ini bertujuan mengevaluasi permasalahan kedua, yakni tingkat pemahaman mahasiswa dan peneliti psikologi di Indonesia mengenai statistik inferensial. Secara lebih spesifik, peneliti hendak mengevaluasi sejauh mana pemahaman mahasiswa dan peneliti psikologi mengenai konsep-konsep statistika yang erat kaitannya dengan penentuan besaran sampel, yakni besaran efek, kekuatan uji statistik, dan kriteria signifikansi. Selain itu, terlepas dari pemahaman mereka terhadap konsep-konsep tersebut, Peneliti hendak mengevaluasi kebiasaan mereka dalam menentukan besaran sampel dalam penelitian. Melalui penelitian ini, Peneliti berharap mampu memberikan gambaran keadaan pemahaman statistika peneliti Psikologi, serta mengajukan saran yang bisa membantu untuk menghindari potensi krisis replikasi Psikologi di Indonesia.

Kriteria signifikansi, kekuatan uji statistik, dan besaran efek

Dalam sebagian besar penelitian psikologi, peneliti tertarik menguji apakah terdapat perbedaan antara dua atau lebih kelompok. Misalnya, sebuah penelitian hendak menguji efek sebuah terapi psikologis terhadap *well-being* klien. Peneliti memiliki hipotesis bahwa terapi psikologis tersebut memiliki efek positif terhadap *well-being* klien. Secara operasional, peneliti tersebut hendak menguji apakah skor *well-being* kelompok eksperimen (yang diberi terapi) lebih tinggi dari skor *well-being* kelompok kontrol.

Kerangka statistik inferensial tradisional tidak memungkinkan pengujian seperti di atas (baca bagian "The Fisherian Legacy", Cohen, 1990). Sebaliknya, statistik inferensial tradisional memungkinkan uji statistik yang disebut *Null Hypothesis Significance Testing* (NHST). Dalam NHST, peneliti membangun hipotesis nihil, yang umumnya merupakan negasi dari hipotesis yang hendak diuji peneliti. Menggunakan contoh terapi *well-being* di atas, formulasi hipotesis nihil dapat berupa: skor kesejahteraan hidup dari kelompok eksperimen sama atau lebih rendah dari kelompok kontrol (H_0: $WB_e \leq WB_k$). Peneliti kemudian melakukan pengambilan data (msl. mengukur skor *well-being* pasca terapi), dan menghitung probabilitas jangka panjang data tersebut muncul di bawah hipotesis nihil, misalnya menggunakan uji *t* dua kelompok independen (*independent samples t-test*). Jika probabilitas jangka panjang data yang diobservasi muncul di bawah hipotesis nihil sangat rendah (misal di bawah 5%), peneliti menyimpulkan bahwa hipotesis nihil sangat kecil kemungkinannya untuk benar. Dengan kata lain, sangat kecil kemungkinan bahwa tidak ada perbedaan antara kelompok kontrol dan kelompok eksperimen. Karena sangat kecil kemungkinan bahwa hipotesis nihil benar, maka peneliti menolak hipotesis nihil, dan menyimpulkan bahwa terapinya memberikan efek positif terhadap kesejahteraan hidup.

NHST memiliki beberapa parameter untuk mencegah terjadinya pengambilan keputusan yang keliru, yakni kriteria

signifikansi, besaran efek, dan kekuatan uji statistik.

Kriteria signifikansi

Dalam NHST, kesalahan Tipe I dilambangkan dengan simbol α (alfa), dan merupakan probabilitas jangka panjang sebuah penelitian menolak hipotesis nihil, ketika hipotesis nihil benar. Sebagian besar penelitian Psikologi dan humaniora memberikan toleransi 5% (α = 0.05) terhadap terjadinya kesalahan Tipe I atau atau biasa disebut *false positives* (Cohen, 1988, 1992a, 1992b). Artinya, dalam observasi (baca: pengambilan data) yang dilakukan berkali-kali dengan batas tidak terhingga, terdapat kemungkinan 5% atau kurang bahwa efek yang ditemukan dalam observasi sebenarnya tidak ada. Dalam praktiknya, toleransi terhadap kesalahan Tipe I dilambangkan dengan *p* (baca: nilai *p*), sehingga nilai *p* di bawah 0,05 (di bawah batas toleransi) dianggap sebagai temuan yang signifikan dan sebaliknya.

Sebagai ilustrasi, bayangkan sebuah skenario di mana Anda diminta mengambil sebutir intan dari sebuah kaleng yang berisi intan dan kerikil tanpa melihat isi kaleng. Tergantung kemampuan Anda membedakan tekstur antara intan dan kerikil, Anda melakukan kesalahan Tipe I ketika mengambil kerikil dan menebaknya sebagai intan. Dalam skenario terapi yang dijelaskan sebelumnya, Anda menyimpulkan bahwa terapi Anda memiliki efek positif, padahal terapi tersebut tidak memiliki efek positif.

Kekuatan uji statistik/statistical power

Dalam NHST, kesalahan Tipe II dilambangkan dengan simbol β (beta), dan merupakan probabilitas jangka panjang sebuah penelitian gagal menolak hipotesis nihil, ketika hipotesis nihil tidak benar. Kekuatan uji statistik (*power*) dalam statistik inferensial tradisional merupakan kontrol terhadap kesalahan Tipe II atau disebut juga sebagai *false negatives* (Cohen, 1988, 1992a, 1992b) (1- β), sebagaimana nilai *p* merupakan kontrol terhadap kesalahan Tipe I. Sebagian besar penelitian Psikologi dan humaniora memberikan toleransi 20% terjadinya kesalahan Tipe II, sehingga penelitian-penelitian tersebut memiliki kekuatan uji statistik sebesar 80% (Cohen, 1990). Artinya, dalam observasi (baca: pengambilan data) yang dilakukan berkali-kali dengan batas tidak terhingga, terdapat kemungkinan 80% atau lebih untuk menyimpulkan bahwa suatu efek tidak ada, ketika efek tersebut memang tidak ada.

Sebagai ilustrasi, dalam skenario intan dan kerikil, Anda mengambil intan, dan menebaknya sebagai kerikil. Dalam skenario terapi psikologis, Anda baru saja menyimpulkan bahwa terapi Anda tidak memiliki efek positif, padahal terapi tersebut memiliki efek positif.

Besaran efek

Besaran efek (*effect size*) menunjukkan perbedaan terstandar antara nilai observasi dari kelompok kontrol dan eksperimen (Durlak, 2009; Lakens, 2013). Dalam penelitian, peneliti tidak hanya tertarik apakah perbedaan antara kelompok kontrol dan eksperimen lebih besar dari nol (kontrol terhadap kesalahan Tipe I), namun juga seberapa besar perbedaan antara kelompok kontrol dan kelompok eksperimen. Tergantung dari pertanyaan penelitian, perbedaan terstandar antara kelompok kontrol dan eksperimen dapat diukur menggunakan parameter yang berbeda. Dalam skenario intan dan kerikil misalnya, besaran efek dapat diukur berdasarkan perbedaan rasio frekuensi antara kelompok kontrol dan kelompok eksperimen. Dalam skenario terapi

psikologis, perbedaan dapat diukur dari rerata skor kesejahteraan hidup antara kelompok eksperimen (diberi terapi) dan kelompok kontrol (tidak diberi terapi).

Besaran efek merupakan satuan standar. Artinya, besaran efek dapat dibandingkan antar beberapa skala yang berbeda dan dapat dibandingkan antar beberapa penelitian dengan besaran sampel yang berbeda-beda. Sebagai ilustrasi, bayangkan dua penelitian yang bertujuan mengukur efek terapi psikologis terhadap kesejahteraan hidup. Penelitian pertama menggunakan pengukuran kesejahteraan hidup dengan skala 1-5 sedangkan penelitian kedua menggunakan pengukuran kesejahteraan hidup dengan skala 1-100. Penelitian pertama melaporkan rerata kelompok kontrol = 3 (*SD* = 0,07) dan rerata kelompok eksperimen = 3.5 (*SD*=0,05). Penelitian kedua melaporkan rerata kelompok kontrol = 62 (*SD* = 19) dan rerata kelompok eksperimen = 68 (*SD* = 25). Besaran efek diperoleh melalui penghitungan berikut:

$$d = \frac{\acute{X}_1 - \acute{X}_2}{S_{within}} \quad (1)$$

Pada penghitungan (1), *d* merupakan perbedaan rerata standar antar dua kelompok dengan satuan standar deviasi, sedangkan $\acute{X}_1$ dan $\acute{X}_2$ merupakan rerata skor dari kelompok eksperimen dan kelompok kontrol. Penyebut S_{within} merupakan simpangan baku dalam kelompok, yang dihitung sebagai berikut:

$$S_{within} = \sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_2+n_2-2}} \quad (2)$$

Pada penghitungan (2), S_1 dan S_2 merupakan standar kesalahan di penelitian pertama dan kedua, sedangkan n_1 dan n_2 merupakan besaran sampel di penelitian pertama dan kedua. Berdasarkan hasil penghitungan, dapat disimpulkan bahwa penelitian pertama menemukan efek terapi yang lebih besar, dengan perbedaan antara kelompok kontrol dan eksperimen sebesar 0,82 standar deviasi dibandingkan penelitian kedua, *d* = 0,27 atau 0,27 standar deviasi. Kesimpulan tersebut terlepas dari ada tidaknya kontrol terhadap kesalahan Tipe I dan II yang digunakan dalam penelitian-penelitian tersebut.

Besaran sampel

Besaran sampel dalam NHST dapat dihitung dengan memanfaatkan dinamika relasi antara parameter-parameter yang telah dijabarkan di paragraf-paragraf sebelumnya. Penghitungan besaran sampel untuk uji *t* dua kelompok independen adalah sebagai berikut.

$$n_i = 2\left(\frac{Z_{1-\alpha/2}+Z_{1-\beta}}{d}\right)^2 \quad (3)$$

di mana n_i merupakan besaran sampel tiap-tiap kelompok, *z* merupakan skor *z* dari distribusi normal di bawah probabilitas 1-α/2 dan 1-β, dan *d* merupakan besaran efek.

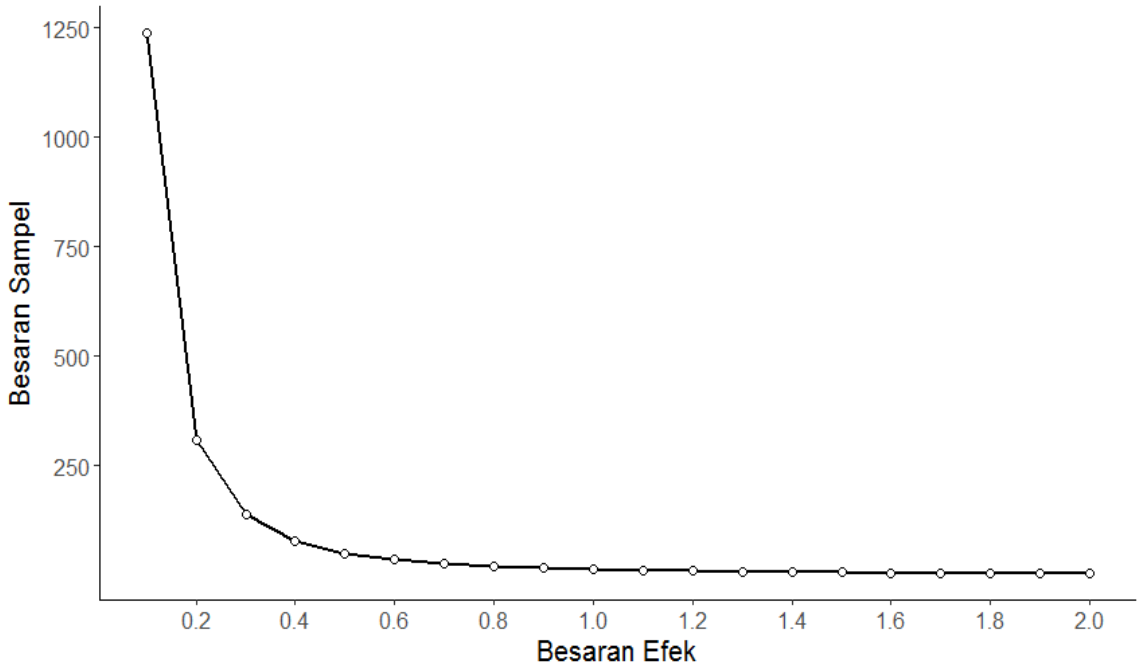
Sebagai ilustrasi, jika kita menentukan parameter kriteria signifikansi, α = 0,05, *two-tailed* dan *power*, 1-β = 0,80 dan mengharapkan besaran efek sebesar 0,5 standar deviasi maka dapat dihitung $Z_{1-\alpha/2}$ = 1,960 dan $Z_{1-\beta}$ = 0,842 (menggunakan tabel distribusi normal) sehingga diperoleh besaran sampel 2*((1,960+0,842)/0,5)^2 = ~ 63 partisipan per kelompok. Formula yang sama dapat digunakan untuk menghitung kekuatan uji statistik, besaran efek, atau kriteria signifikansi jika tiga parameter lainnya diketahui. Sebagai ilustrasi, sebuah penelitian dengan jumlah sampel 30 per kelompok dengan besaran efek sebesar 0,5 dan kriteria signifikansi 0,05 memiliki *power* sebesar 0,478. Artinya, penelitian dengan sampel 30 partisipan memiliki taraf kesalahan Tipe II sebesar 52,2%. Dengan kata lain, terdapat 52,2% kemungkinan dalam penelitian ini untuk menolak

hipotesis nihil, padahal hipotesis nihil tersebut benar! Taraf kesalahan sebesar 52,2% ini melebihi toleransi kesalahan Tipe II sebesar 20% dalam sebagian besar penelitian Psikologi dan humaniora.

Menggunakan penghitungan (3), dengan asumsi kriteria signifikansi α dan kekuatan uji statistik 1-β konstan, dapat dibuat grafik relasi antara besaran efek dan besaran sampel dalam analisis *t-test* dua kelompok independen sebagai berikut (Gambar 1).

Terdapat beberapa hal yang perlu diperhatikan dari Gambar 1. Pertama, ketika kesalahan Tipe I dan II dikontrol, besaran efek memiliki relasi negatif dengan besaran sampel. Untuk mendeteksi besaran efek yang kecil, dibutuhkan sampel yang lebih banyak dan sebaliknya. Sebagai analogi yang sederhana, penggunaan mata telanjang cukup untuk melihat benda-benda langit yang dekat seperti bulan (besaran efek yang besar), namun untuk benda-benda langit yang jauh seperti bintang (besaran efek yang kecil), dibutuhkan alat bantu lihat yang lebih sensitif (besaran sampel yang besar). Kedua, besaran efek memiliki relasi eksponensial dengan besaran sampel. Ketika besaran efek yang diharapkan kecil, maka besaran sampel yang dibutuhkan untuk mendeteksi suatu efek bertambah secara eksponensial menjadi sangat besar dan sebaliknya.

Dalam banyak penelitian psikologi dan humaniora, peneliti mengontrol taraf kesalahan Tipe I, namun tidak mengontrol taraf kesalahan Tipe II dan tidak mempertimbangkan besaran efek dalam pengambilan keputusan (Bakker, van Dijk, & Wicherts, 2012; Gigerenzer, 2004). Ketika taraf kesalahan Tipe II tidak dikontrol, besaran efek yang terhitung biasanya merupakan overestimasi, dan ketika besaran efek ini digunakan untuk menghitung kekuatan uji statistik dalam penelitian replikasi, hasil penelitian sebagian besar tidak mereplikasi temuan sebelumnya (Button *et al*., 2013; Ioannidis,



Gambar 1. Relasi antara besaran efek dan besaran sampel dengan α = .05 dan 1-β = 20

2008; Open Science Collaboration, 2015). Lebih lanjut, penelitian juga menunjukkan bahwa kebiasaan tidak mengontrol taraf kesalahan Tipe II dan besaran efek berkaitan erat dengan kekurangpahaman terhadap kekuatan uji statistik dan pengukuran besaran efek (Button *et al*., 2013). Sebagian besar buku teks statistika yang digunakan dalam psikologi hanya sedikit atau sama sekali tidak menyinggung kedua topik tersebut (Gigerenzer, 2004). Kondisi tersebut semakin serius seiring dengan menguatnya anggapan bahwa peneliti dan editor jurnal psikologi lebih mengutamakan hasil penelitian yang berbeda secara signifikan (baca: *p-value* kecil) untuk dipublikasikan (Gigerenzer, 2004). Hal ini semakin menjadi pendorong bagi peneliti untuk memprioritaskan hasil penelitian yang menunjukkan signifikansi untuk dilaporkan dalam artikel. Dengan pertimbangan bahwa mendapatkan hasil penelitian yang positif (nilai $p < 0,05$) penting dalam publikasi ilmiah, buku teks dan kelas-kelas statistik mempertahankan praktik NHST sebagai model uji statistik yang tepat dalam penelitian Psikologi (Wasserstein & Lazar, 2016). Padahal, simulasi menunjukkan bahwa temuan-temuan dalam publikasi ilmiah Psikologi melaporkan perbedaan yang signifikan dengan frekuensi yang jauh lebih tinggi dari ideal (Bakker *et al*., 2012).

Metode

Partisipan penelitian

Partisipan penelitian merupakan mahasiswa, peneliti, dan dosen Psikologi di Indonesia yang diundang mengisi survei melalui media sosial dan/atau melalui kontak pribadi. Pengisian kuesioner dilakukan secara daring menggunakan jasa yang dikelola Qualtrics (www.qualtrics.com). Sebanyak N = 354 partisipan menerima undangan peneliti mengisi kuesioner. Dari jumlah tersebut, sebanyak N = 151 partisipan mengisi kuesioner secara lengkap.

Survei

Survei yang digunakan dalam penelitian ini memiliki tiga bagian. Bagian pertama berisi pertanyaan mengenai identitas partisipan; bagian kedua berisi pertanyaan mengenai familiaritas dengan istilah-istilah statistika terkait perhitungan besaran sampel; bagian ketiga berisi pertanyaan mengenai kebiasaan partisipan menentukan besaran sampel. Pada bagian kedua, khususnya pada pertanyaan mengenai makna istilah-istilah statistik, peneliti meminta partisipan menilai kebenaran dari 4-5 pernyataan (Gigerenzer, 2004., Hoekstra *et al*., 2014). Dari 4-5 pernyataan tersebut, hanya terdapat satu pernyataan yang benar, namun partisipan memiliki kesempatan menilai seluruh pernyataan sebagai benar atau salah.

Analisis data dan simulasi

Data dianalisis dalam dua bagian. Pada bagian pertama, peneliti menjabarkan temuan deskriptif menggunakan tabel kontingensi (*contingency table*) berisi frekuensi respon partisipan, rerata, dan simpangan baku untuk tiap pertanyaan yang peneliti ajukan. Selanjutnya, peneliti menguji secara inferensial apakah partisipan dengan tingkat pendidikan yang berbeda memiliki perbedaan preferensi dalam mendukung pernyataan yang tepat. Pengujian tersebut menggunakan *chi-square k* kelompok independen. Pada bagian kedua, peneliti melakukan simulasi perolehan kekuatan uji statistik untuk besaran efek, berdasarkan respon partisipan mengenai besaran sampel ideal. Untuk menyederhanakan, simulasi tersebut hanya dilakukan untuk uji *t* dua kelompok.

Seluruh perhitungan dilakukan melalui R (R Core team, 2015). Simulasi dibuat menggunakan paket pwr (Champely, 2009). Grafik dilukis menggunakan paket ggplot2 (Wickham, 2010).

Hasil

Demografi partisipan

Seperti terlihat di Tabel 1, lebih dari tiga per empat (75,49%) partisipan mengidentifikasikan diri sebagai mahasiswa Psikologi. Selanjutnya, sebagian besar (40,40%) partisipan melaporkan bahwa saat ini mereka tengah menempuh pendidikan S1 Psikologi, dan sudah lulus beberapa mata kuliah statistika (Tabel 2).

Sebagian besar partisipan jarang (31-50% dari penelitian) menerapkan pengetahuan mengenai statistika dalam penelitian (misal: dalam menulis/ membimbing skripsi, blog pribadi, dan artikel ilmiah; Tabel 3).

Pemahaman statistika

Familiaritas terhadap istilah-istilah dalam statistika. Sebagian besar partisipan lebih familiar dengan nilai p dibandingkan dengan besaran efek (*effect size*) dan kekuatan uji (*power*) dalam statistika. Istilah besaran efek juga lebih familiar dibandingkan kekuatan uji (Tabel 4).

Tabel 1.
Latar Belakang Partisipan

	Mahasiswa	Dosen (Stat)	Dosen (Bukan Stat)	Dosen (MetPen)	Dosen (Bukan MetPen)	Peneliti (Psi)	Peneliti (Bukan Psi)
Jumlah	114	4	13	11	11	21	19
%	75.49	2.65	8.61	7.29	7.29	13.91	12.58

Tabel 2
Pendidikan Partisipan

	Menempuh S1	Menempuh S1 (lulus Stat)	S1	menempuh Profesi	menempuh Sains	Psikolog	Psikolog + Magister	Magister Sains	Menempuh Doktor	Doktor
Jumlah	15	61	18	8	13	4	5	18	7	2
%	9.93	40.40	11.92	5.30	8.61	2.65	3.31	11.92	4.64	1.32

Tabel 3
Pengalaman Responden dalam Aplikasi Ilmu-ilmu Statistika

	< 15%	15-30%	31-50%	51-70%	71-85%	> 85%
Jumlah	28	36	38	26	8	15
%	18.54	23.84	25.17	17.22	5.30	9.93

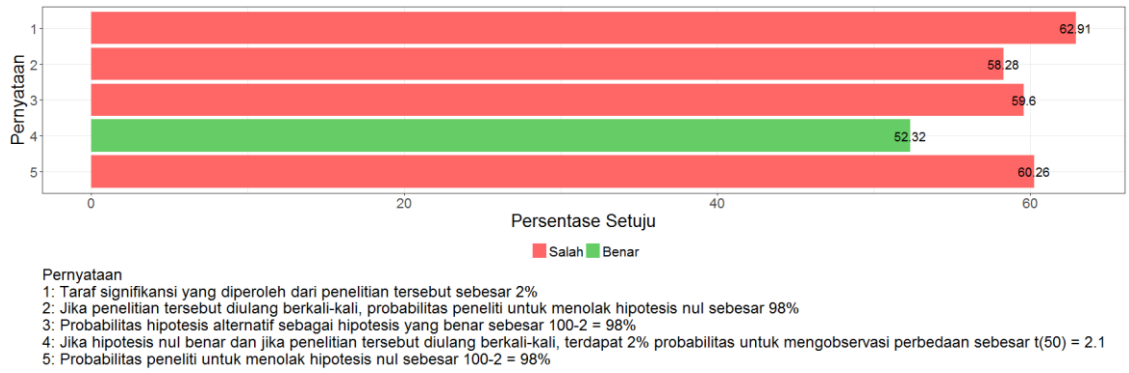
Tabel 4.
Familiaritas dengan Istilah-istilah Statistika

	Sangat tidak familiar	Tidak familiar	Familiar	Sangat familiar	Rerata	Simpangan Baku
Nilai p	6	18	70	57	3.18	0.79
Besaran efek	17	55	62	17	2.52	0.84
Kekuatan uji statistik	19	74	44	14	2.35	0.82

Pemahaman mengenai nilai p. Meski mengaku familiar dengan nilai p, hampir setengah partisipan tidak mendukung pernyataan yang tepat mengenai nilai p, yakni pernyataan (4): "Jika hipotesis nihil benar dan jika penelitian tersebut diulang berkali-kali, terdapat 2% probabilitas untuk mengobservasi perbedaan sebesar $t(50) = 2.1$." Pernyataan ini benar karena selain merujuk pada probabilitas mengobservasi data mengingat H_0, $p(\text{data}|H_0)$, secara spesifik dijelaskan juga bahwa probabilitas tersebut diperoleh dalam pengambilan data yang diulang berkali-kali. Pernyataan tersebut merupakan pernyataan yang paling sedikit didukung dari lima pernyataan yang diberikan sebagai pilihan (Gambar 2). Lebih lanjut, hanya empat partisipan (2,65%) yang mendukung pernyataan ini dan tidak mendukung pernyataan-pernyataan yang lain. Hanya empat partisipan yang memiliki pemahaman yang utuh mengenai *nilai p*.

Pernyataan selain pernyataan (4) merupakan pernyataan yang salah. Pernyataan (1) tidak secara spesifik menjelaskan apa yang dimaksud dengan signifikansi, sedangkan pernyataan (3) mengindikasikan probabilitas H_1 berdasarkan data, $p(H_1|\text{data})$. Pernyataan (5) merujuk pada probabilitas untuk menolak H_0 berdasarkan data, $p(H_0|\text{data})$, dan pernyataan (2), meskipun secara spesifik mengindikasikan pengambilan data berulang, tetap merujuk pada probabilitas menolak H_0 berdasarkan data $p(H_0|\text{data})$.

Pernyataan "Taraf signifikansi yang diperoleh dari penelitian tersebut sebesar 2%" merupakan pernyataan yang paling banyak didukung sebagai pernyataan yang benar. Lebih lanjut, partisipan dengan level pendidikan yang berbeda tidak menunjukkan perbedaan preferensi dalam mendukung pernyataan yang tepat, $\chi^2(9) = 8{,}52$, $p = 0{,}482$.



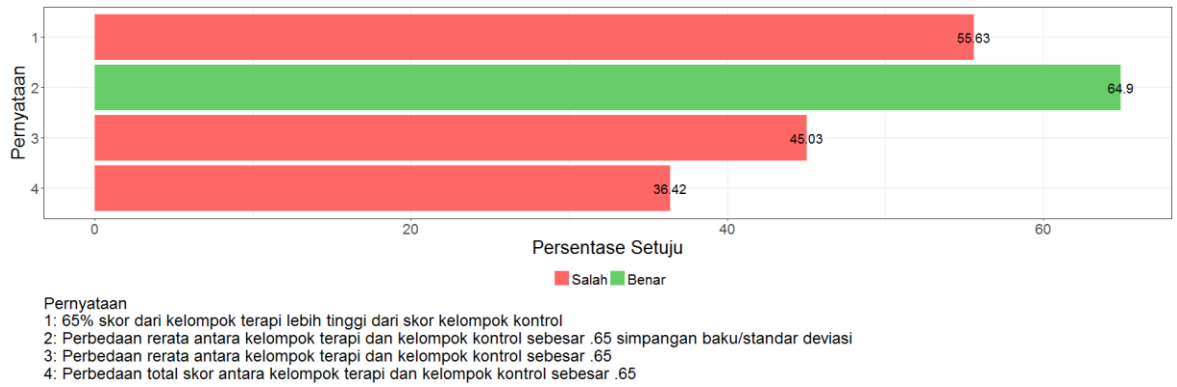
Gambar 2. Sebaran dukungan partisipan terhadap pernyataan mengenai nilai p.
Keterangan: Warna balok berkorespondensi dengan kebenaran pernyataan dengan hijau = benar dan merah = salah.

Pemahaman mengenai besaran efek. Pernyataan (2) "Perbedaan rerata antara kelompok terapi dan kelompok kontrol sebesar .65 simpangan baku." merupakan pernyataan yang tepat dan pernyataan yang paling banyak didukung sebagai pernyataan yang benar (Gambar 3). Pernyataan (2) menggunakan parameter yang benar (rerata) dan secara spesifik merujuk pada skor standar yang merupakan ciri khas dari besaran efek. Namun, hanya 34 partisipan (22,52%) yang mendukung pernyataan ini dan tidak mendukung pernyataan-pernyataan yang lain. Hanya 34 partisipan yang memiliki pemahaman yang utuh mengenai besaran efek.

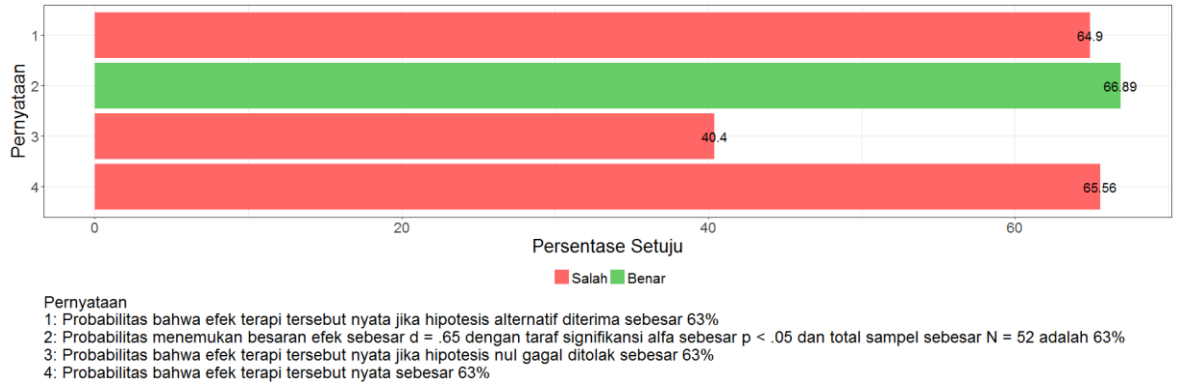
Pernyataan selain pernyataan (2) merupakan pernyataan yang salah. Pernyataan (1) merujuk pada proporsi skor, bukan rerata skor sebagaimana umumnya digunakan sebagai dalam uji statistik parametrik, begitu juga dengan pernyataan (4) yang merujuk pada total skor. Pernyataan (3) tidak secara spesifik menjelaskan standarisasi yang merupakan ciri khas besaran efek.

Tidak terdapat perbedaan preferensi dalam mendukung pernyataan-pernyataan mengenai besaran efek antara partisipan dengan level pendidikan yang berbeda, $\chi^2(9) = 5{,}31$, $p = 0{,}802$.

Pemahaman mengenai kekuatan uji statistik. Pernyataan (2): "Probabilitas menemukan besaran efek sebesar $d = 0{,}65$ dengan taraf signifikansi alfa sebesar $p < 0{,}05$ dan total sampel $N = 52$ sebesar 63%." merupakan pernyataan yang tepat dan pernyataan yang paling banyak didukung sebagai pernyataan yang benar (Gambar 4). Pernyataan ini merupakan pernyataan yang tepat karena merujuk pada probabilitas untuk gagal menolak hipotesis nihil yang salah berdasarkan parameter besaran efek, kriteria signifikansi alfa, dan besaran sampel $p(\text{data}|H_0)$. Namun, hanya 15 partisipan (9,93%) yang mendukung pernyataan ini dan tidak mendukung pernyataan-pernyataan yang lain. Hanya sembilan partisipan yang memiliki pemahaman yang utuh mengenai kekuatan uji statistik.



Gambar 3. Sebaran dukungan partisipan terhadap pernyataan mengenai besaran efek.
Keterangan: Warna balok berkorespondensi dengan kebenaran pernyataan dengan hijau = benar dan merah = salah.



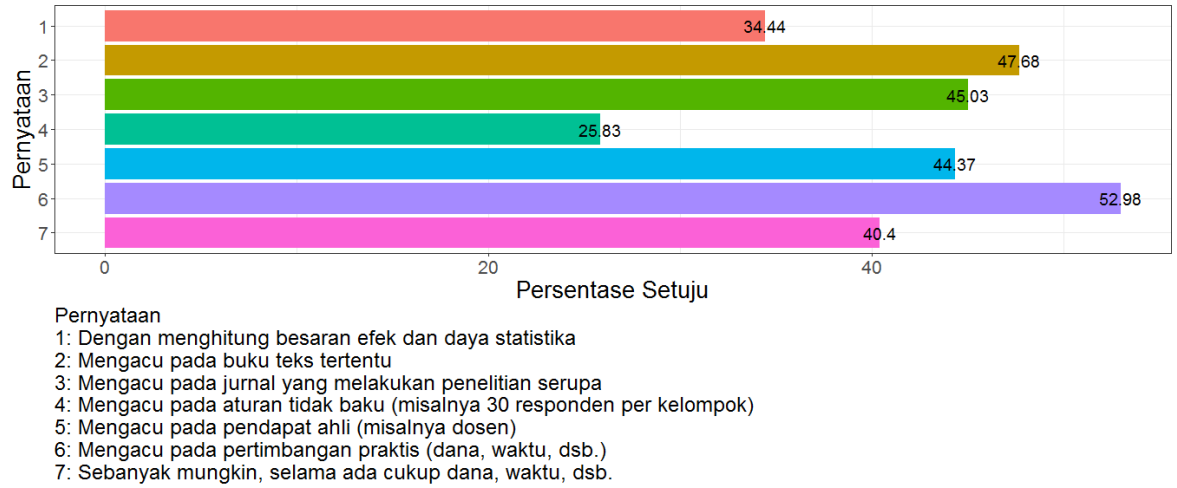
Gambar 4. Sebaran dukungan partisipan terhadap pernyataan mengenai kekuatan uji statistik. Keterangan: Warna balok berkorespondensi dengan kebenaran pernyataan dengan hijau = benar dan merah = salah.

Pernyataan selain pernyataan (2) merupakan pernyataan yang salah. Pernyataan (4) tidak secara spesifik menjelaskan apa yang dimaksud sebagai efek yang "nyata." Pernyataan (1) merujuk pada probabilitas H_1 berdasarkan besaran efek $p(H_1 | \text{data})$ dan pernyataan (4) merujuk pada probabilitas H_0 berdasarkan besaran efek $p(H_0 | \text{data})$.

Tidak terdapat perbedaan preferensi dalam mendukung pernyataan-pernyataan mengenai besaran efek antara partisipan dengan level pendidikan yang berbeda, $\chi^2(9) = 5{,}79$, $p = 0{,}761$.

Besaran sampel penelitian

Lebih dari setengah (52,98%) partisipan mengacu pada pertimbangan praktis dalam menentukan besaran sampel. Praktik menghitung besaran sampel berdasarkan besaran efek dan *power* sebagai metode ideal menempati urutan kedua dari bawah (34,4%) setelah mengikuti aturan tidak baku (25,83%; Gambar 5).



Gambar 5. Sebaran dukungan partisipan terhadap pernyataan mengenai praktik menentukan besaran sampel.

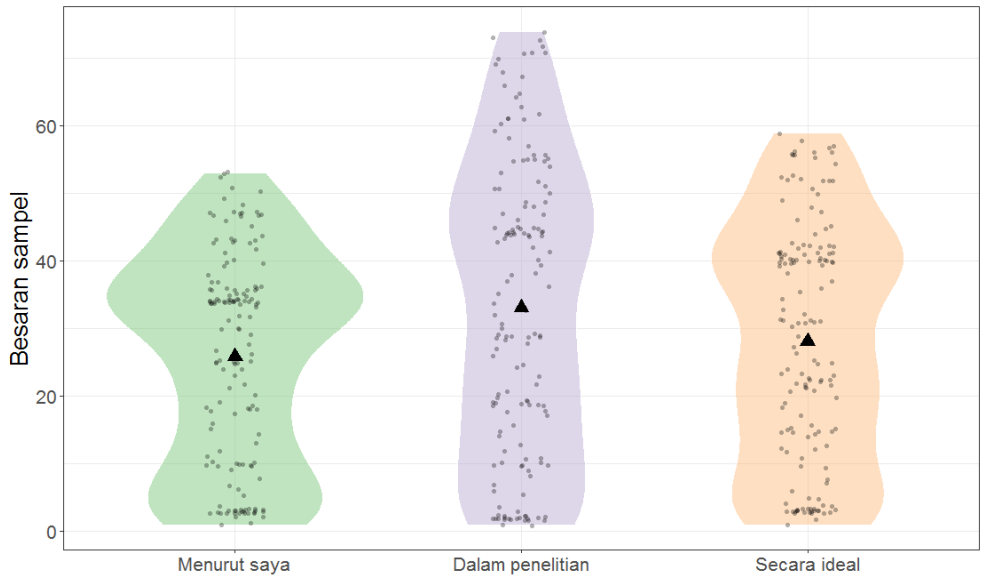
Gambar 5 menunjukkan sebaran respon partisipan mengenai jumlah sampel (1) yang ideal menurut partisipan, (2) dalam sebagian besar penelitian, dan (3) secara ideal sejauh pengetahuan partisipan. Rata-rata jumlah sampel penelitian yang ideal menurut partisipan adalah 26 partisipan per kelompok (M = 25,91, SD = 15,54). Dalam sebagian besar penelitian yang dibaca partisipan, jumlah sampel yang dicantumkan adalah 33 partisipan per kelompok (M = 33,1; SD = 21,80). Sedangkan secara ideal, sejauh pengetahuan partisipan, jumlah sampel dalam penelitian adalah 28 partisipan per kelompok (M = 28,15; SD = 17,54).

Gambar 6 menunjukkan sebaran respon partisipan. Untuk skenario "menurut saya," besaran sampel berkisar antara 1-53 responden per kelompok, untuk skenario "dalam penelitian," besaran sampel berkisar antara 1-74 responden per kelompok, dan untuk skenario "secara ideal," besaran sampel berkisar antara 1-59 responden per kelompok.

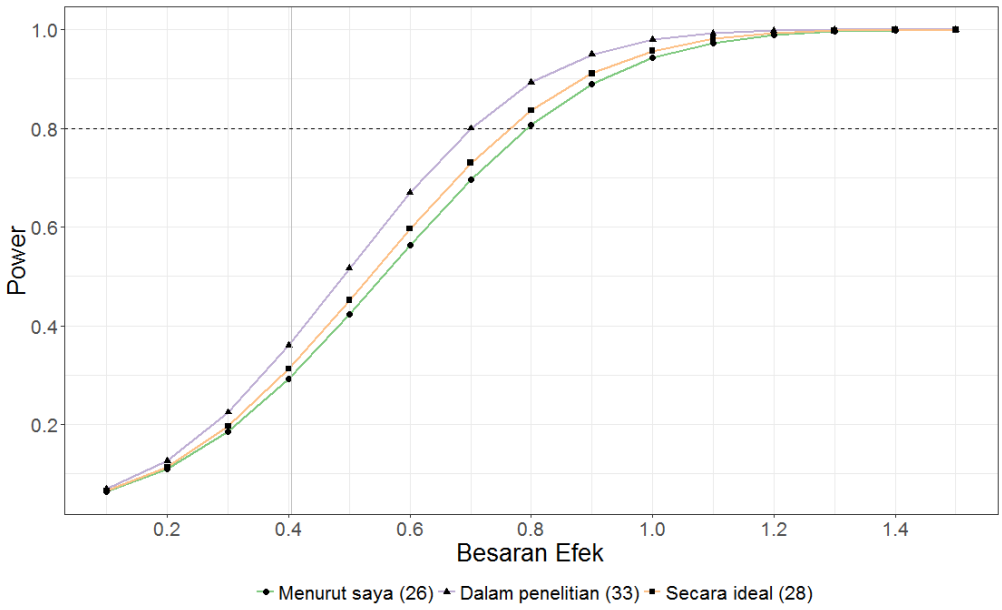
Simulasi besaran sampel

Menggunakan informasi di atas dan penghitungan (3), dapat dibuat simulasi kekuatan uji statistik yang diperoleh menggunakan besaran sampel yang diajukan responden untuk besaran efek yang berbeda-beda. Hasil simulasi dapat dilihat di Gambar 7.

Pada skenario di mana partisipan menentukan besaran sampel berdasarkan pendapat pribadinya ("menurut saya"), hasil uji t dua kelompok independen baru mencapai kekuatan uji statistik 80% ketika fenomena yang diuji memiliki besaran efek d = 0,79 atau lebih tinggi. Mengingat rerata besaran efek dalam penelitian Psikologi yang dilaporkan *Open Science Collaboration* adalah d = 0,403, besaran sampel ini ("menurut saya") hanya memiliki kekuatan uji statistik sebesar 29,67%. Artinya, terdapat kemungkinan 70,33% peneliti gagal menolak hipotesis nihil yang salah, atau menyimpulkan bahwa tidak ada efek ketika efek tersebut sebenarnya ada.



Gambar 6. Plot biola mengenai besaran sampel yang ideal.
Keterangan: Simbol segitiga menunjukkan rerata jawaban per pernyataan; titik-titik hitam menunjukkan sebaran respon partisipan.



Gambar 7. Simulasi kekuatan uji statistik menggunakan besaran sampel menurut partisipan pada kriteria signifikansi alfa = 0,05.
Keterangan: Garis titik-titik horizontal menunjukkan *power* = 0,80 dan garis abu-abu menunjukkan rerata besaran efek dalam penelitian Psikologi yang dilaporkan oleh *Open Science Collaboration*, d = 0,403.

Selanjutnya, pada skenario di mana partisipan menentukan besaran sampel berdasarkan penelitian yang mereka tahu ("berdasarkan penelitian"), hasil uji t dua kelompok independen mencapai kekuatan uji statistik 80% ketika mendeteksi besaran efek d = 0,70 atau lebih tinggi. Akibatnya, ketika digunakan untuk mendeteksi besaran efek dalam sebagian besar penelitian Psikologi yang dilaporkan *Open Science Collaboration*, besaran sampel ini ("berdasarkan penelitian") hanya memiliki kekuatan uji statistik sebesar 36,43%. Artinya, terdapat kemungkinan 63,57% peneliti gagal menolak hipotesis nihil yang salah, atau menyimpulkan bahwa tidak ada efek ketika efek tersebut sebenarnya ada.

Terakhir, pada skenario di mana partisipan menentukan besaran sampel berdasarkan pendapat mereka mengenai besaran sampel ideal ("secara ideal"), hasil uji t dua kelompok independen mencapai kekuatan uji statistik 80% ketika mendeteksi besaran efek d = 0,76 atau lebih tinggi. Akibatnya, untuk mendeteksi besaran efek dalam sebagian besar penelitian Psikologi yang dilaporkan *Open Science Collaboration*, besaran sampel ini ("secara ideal") hanya memiliki kekuatan uji statistik sebesar 31,63%. Artinya, terdapat kemungkinan 68,37% peneliti gagal menolak hipotesis nihil yang salah, atau menyimpulkan bahwa tidak ada efek ketika efek tersebut sebenarnya ada.

Mempertimbangkan ketiga skenario di atas menghasilkan proyeksi kekuatan uji statistik yang relatif rendah, peneliti membuat satu simulasi lagi menggunakan besaran sampel maksimal yang dilaporkan partisipan dalam skenario "dalam penelitian," n = 74. Menggunakan besaran sampel ini, diperoleh kekuatan uji statistik 80% untuk mendeteksi besaran efek d = 0,46 atau lebih tinggi, dan kekuatan uji statistik 68,28% untuk mendeteksi besaran efek dalam sebagian besar penelitian psikologi.

Diskusi

Dalam penelitian ini, partisipan diminta melaporkan familiaritasnya terhadap berbagai istilah statistika, menjawab pertanyaan mengenai konsep-konsep statistika, dan melaporkan besaran sampel yang menurut mereka ideal dalam penelitian. Secara umum, hasil penelitian menunjukkan bahwa mayoritas partisipan cukup familiar terhadap istilah-istilah statistika, namun memiliki pemahaman yang keliru mengenai istilah-istilah tersebut. Lebih lanjut, besaran sampel yang dilaporkan partisipan hanya memiliki kekuatan uji statistik sebesar 29,67-36,43% untuk menolak hipotesis nihil yang salah. Artinya, jika pemahaman partisipan mengenai besaran sampel yang dilaporkan mencerminkan besaran sampel yang mereka gunakan dalam penelitian, pada banyak kasus partisipan akan gagal mendeteksi fenomena psikologis yang memiliki besaran efek kecil (*Cohen's d* < .8).

Respon partisipan mengenai istilah-istilah statistika menunjukkan dua paradoks yang menarik. Pertama, sebagian besar partisipan mengaku familiar terhadap nilai *p*, namun sebagian besar memiliki pemahaman yang keliru mengenai nilai *p*. Ini bukan temuan yang baru. Berbagai penelitian sebelumnya juga telah menunjukkan bahwa mahasiswa dan peneliti Psikologi dari berbagai universitas, baik universitas yang memiliki reputasi yang baik hingga universitas yang tidak terlalu dikenal memiliki kesulitan yang sama dalam menginterpretasikan nilai *p* (Badenes-Ribera *et al*., 2016, 2015; Bakker *et al*., 2012; Gigerenzer, 2004). Asosiasi Statistika Amerika Serikat (ASA) baru-baru ini bahkan mempublikasikan panduan mengenai apa itu nilai *p* dan bagaimana menginterpretasikan nilai *p* dengan baik untuk meningkatkan pemahaman mahasiswa dan peneliti Psikologi karena banyaknya penyalahgunaan dan penyalahartian nilai *p* dalam publikasi ilmiah (Wasserstein & Lazar, 2016).

Salah satu alasan mengapa nilai *p* sulit diinterpretasikan adalah karena interpretasi nilai *p* membutuhkan kemampuan berpikir logis yang kuat. Seperti dibahas sebelumnya, dalam banyak penelitian, peneliti berharap mengevaluasi hipotesis berdasarkan data $p(H_1|\text{data})$, sedangkan nilai *p* menginformasikan probabilitas mengobservasi data dengan asumsi bahwa negasi dari hipotesis yang diajukan peneliti benar, $p(\text{data}|H_0)$. Untuk memahami pernyataan terakhir, peneliti pertama harus paham bahwa $p(H_1|\text{data})$ tidak sama dengan $p(\text{data}|H_1)$. Dengan kata lain, nilai *p* yang kecil tidak menunjukkan bahwa hipotesis alternatif merupakan hipotesis yang kemungkinan besar tepat. Kemudian, peneliti harus paham bahwa $p(\text{data}|H_0)$ merupakan negasi dari $p(\text{data}|H_1)$. Dengan kata lain, nilai *p* yang besar tidak menunjukkan bahwa hipotesis nihil, sebagai negasi dari hipotesis alternatif, merupakan hipotesis yang tepat. Untuk memahami kedua hal di atas, dibutuhkan pemahaman mengenai batasan interpretasi nilai *p*. Secara keseluruhan, nilai *p* hanya menunjukkan probabilitas dalam pengambilan data yang tidak terhingga, untuk mengobservasi data jika hipotesis nihil benar (Gigerenzer, 2004; Wagenmakers, 2007). Ketika nilai *p* sangat kecil, dapat disimpulkan bahwa kecil kemungkinan hipotesis nihil merupakan hipotesis yang benar. Menggunakan falsifikasi, dapat kemudian disimpulkan bahwa negasi dari hipotesis nihil (i.e. hipotesis alternatif) merupakan hipotesis yang benar. Sayangnya, terdapat kecenderungan bagi peneliti psikologi untuk menginterpretasi nilai *p* secara berlebihan, entah karena tuntutan dari jurnal yang mensyaratkan hanya temuan dengan nilai *p*

yang kecil yang layak dipublikasikan (bias positif) atau dengan dalih bahwa peneliti lain dari universitas ternama pun menginterpretasi nilai *p* secara berlebihan (Badenes-Ribera *et al*., 2015).

Dari 151 partisipan, hanya empat partisipan yang menunjukkan pemahaman yang utuh mengenai batasan interpretasi nilai *p* seperti yang dibahas di atas. Artinya, sebagian besar partisipan, baik mahasiswa maupun peneliti dan dosen, memiliki pemahaman yang keliru atau tidak utuh mengenai nilai *p*. Menariknya, tidak terdapat perbedaan yang signifikan antara respon mengenai pemahaman nilai *p* antara mahasiswa S1 dengan dosen dan peneliti. Lagi-lagi, ini bukan temuan yang baru (Badenes-Ribera *et al*., 2016, 2015; Gigerenzer, 2004). Temuan ini dapat diinterpretasikan sebagai mahasiswa mempelajari pemahaman nilai p yang keliru dari dosennya (lih. Gigerenzer, 2004). Lebih lanjut, peneliti dan dosen kemungkinan besar mempelajari pemahaman yang keliru ini di bangku kuliah (Wasserstein & Lazar, 2016), menciptakan siklus kekeliruan yang terus berkelanjutan. Memperparah kondisi ini, buku teks statistika yang digunakan di psikologi seringkali mendeskripsikan interpretasi nilai *p* yang cenderung menyesatkan, seperti "probabilitas bahwa perbedaan yang diobservasi adalah nyata" atau "dalam 95 atau lebih dari 100 pengambilan data, dapat diobservasi perbedaan tersebut" (lih. Gigerenzer, 2004).

Paradoks kedua, sebagian besar partisipan mengaku tidak familiar terhadap besaran efek dan *power,* namun secara mengejutkan, mereka mendukung pernyataan yang tepat mengenai makna dari besaran efek dan *power.* Perlu diingat bahwa dalam soal pilihan ganda seperti yang peneliti gunakan dalam penelitian ini, menjawab pertanyaan dengan tepat hanya merupakan indikasi bahwa partisipan memiliki pemahaman yang tepat (Dufresne, Leonard, & Gerace, 2002), karena partisipan dapat menggunakan petunjuk-petunjuk seperti letak pernyataan dan jumlah kata dalam pernyataan untuk membantu menentukan jawaban yang benar. Attali dan Bar-Hillel, (2003) dalam penelitiannya menunjukkan bahwa pembuat soal memiliki kecenderungan menempatkan jawaban yang salah di pernyataan yang letaknya di tengah. Hal ini dapat digunakan untuk membantu menebak jawaban yang benar. Lebih lanjut, seperti respon terhadap pemahaman mengenai nilai *p*, jumlah partisipan yang mendukung pernyataan yang tepat namun tidak pernyataan yang lain relatif sedikit, yakni 34 partisipan untuk pertanyaan mengenai besaran efek dan 15 partisipan untuk pernyataan mengenai kekuatan uji statistik. Seperti respon terhadap pemahaman mengenai nilai *p*, temuan ini menunjukkan bahwa partisipan tidak memiliki pemahaman yang utuh mengenai besaran efek dan kekuatan uji statistik.

Penelitian ini menyiratkan bahwa kekuatan uji statistik merupakan isu yang telah lama diabaikan dalam penelitian psikologi. Banyak buku teks statistika dalam psikologi berfokus pada pentingnya kontrol terhadap eror Tipe I (α), namun sedikit atau tidak sama sekali membahas kontrol terhadap eror Tipe II atau kekuatan uji statistik dalam uji hipotesis (Gigerenzer, 2004). Tidak seperti pemahaman yang keliru mengenai nilai *p* yang diakibatkan sulitnya menginterpretasikan nilai *p* secara tepat, kekeliruan pemahaman terhadap kekuatan uji statistik dan besaran efek bisa saja diakibatkan oleh ketidaktahuan dan bukan pemahaman yang salah.

Terakhir, besaran sampel yang dilaporkan partisipan (rata-rata antara 26-33 responden per kelompok) tidak cukup

untuk mendeteksi besaran efek dalam penelitian-penelitian psikologi, sebagaimana yang dilaporkan dalam artikel *Open Science Collaboration* (2015). Dalam skenario terbaik, di mana peneliti menggunakan n = 74 per kelompok, kekuatan uji statistik mencapai 80% hanya ketika mendeteksi besaran efek *d* = 0,46 atau lebih tinggi. Padahal, untuk mendeteksi besaran efek dalam sebagian besar penelitian Psikologi sebagaimana dilaporkan dalam artikel oleh *Open Science Collaboration*, yaitu *d* = 0,403, besaran sampel tersebut tidak cukup untuk mencapai kekuatan uji statistik sebesar 80%.

Distribusi respon partisipan mengenai besaran sampel yang mereka gunakan seperti dilihat dalam plot biola (Gambar 6) menunjukkan hal yang menarik, bahwa sebagian besar partisipan beranggapan bahwa besaran sampel dengan kisaran 30-40 responden per kelompok merupakan besaran yang ideal. Temuan ini mengkonfirmasi anekdot Cohen (1990) yang menyatakan bahwa sebagian besar buku teks psikologi mengasumsikan bahwa besaran sampel sebesar $n = 30$ per kelompok merupakan besaran yang ideal. Cohen (1990) membuktikan bahwa besaran sampel ini hanya memiliki kekuatan uji statistik sebesar 47% untuk mendeteksi besaran efek $d = .5$. Terlebih lagi, sebaran respon $n \approx 30$ ini lebih mengelompok dalam skenario "menurut saya" dibandingkan dua skenario "dalam penelitian" dan "secara ideal" yang memiliki distribusi yang lebih seragam (*uniform*) dengan varians yang lebih tinggi. Mengingat bahwa mengacu pada buku teks merupakan salah satu pertimbangan utama dalam menentukan besaran sampel (Gambar 5), temuan ini dapat diinterpretasikan sebagai partisipan memiliki disonansi kognitif antara apa yang mereka pelajari dalam buku teks dengan apa yang mereka baca dalam penelitian ilmiah.

Penelitian ini telah memberikan gambaran bahwa pemahaman konsep statistika yang keliru dan kecenderungan menentukan besaran sampel yang kurang tepat di Indonesia berpotensi menimbulkan krisis replikasi, sama seperti fenomena global pada umumnya. Dengan memahami situasi tersebut, dibutuhkan tindak lanjut yang tangkas dari pemangku kepentingan terkait, mengingat keadaan ini bisa juga terjadi bukan hanya pada peneliti di bidang psikologi, melainkan bidang ilmu lainnya juga. Meskipun penelitian ini memiliki tingkat penyelesaian (*completion rates*) kuesioner yang terbilang rendah (42,65%) dari angka ideal yakni 80% (Pluye & Hong, 2014), penelitian ini terbilang berhasil untuk memberi gambaran awal mengenai kesenjangan dalam penelitian psikologi di Indonesia yang menggunakan statistika inferensial.

Kesimpulan

Mahasiswa dan peneliti psikologi di Indonesia sudah cukup familiar terhadap istilah-istilah statistika, namun memiliki pemahaman yang keliru atau tidak utuh mengenai istilah-istilah tersebut. Akibatnya, mereka memproyeksikan besaran sampel yang kurang dari ideal dalam penelitian. Secara spesifik, besaran sampel yang dilaporkan memiliki kekuatan uji statistik yang rendah, sehingga kemungkinan hasil penelitian mereka dapat direplikasi rendah.

Saran

Mengingat ajakan dari Kementerian Pendidikan Tinggi dan Kementerian Keuangan bagi dosen dan peneliti Indonesia untuk mempublikasikan lebih banyak artikel, dan mempertimbangkan bahwa artikel-artikel yang dipublikasikan dalam jurnal ilmiah secara umum tidak dapat direplikasi karena

bias dan/atau kekuatan uji statistik yang tidak optimal, penting bagi dosen dan peneliti untuk mempelajari kembali konsep-konsep mengenai besaran efek, besaran sampel, dan kekuatan uji statistik, serta memperkenalkannya dalam kelas-kelas statistika. Selain itu, temuan penelitian ini juga menunjukkan perlunya merancang mekanisme aktivitas berbagi pengetahuan melalui forum-forum akademik bidang psikometri dan statistika sebagai wadah meningkatkan kualitas penelitian dalam bidang psikologi.

Kepustakaan

Attali, Y., & Bar-Hillel, M. (2003). Guess where: The position of correct answers in multiple-choice test items as a psychometric variable. *Journal of Educational Measurement*, *40*(2), 109–128. doi: 10.1111/j.1745-3984.2003.tb01099.x

Badenes-Ribera, L., Frias-Navarro, D., Iotti, B., Bonilla-Campos, A., & Longobardi, C. (2016). Misconceptions of the p-value among Chilean and Italian academic psychologists. *Frontiers in Psychology*, *7*(August), 1247. doi: 10.3389/fpsyg.2016.01247

Badenes-Ribera, L., Frías-Navarro, D., Monterde-I-Bort, H., & Pascual-Soler, M. (2015). Interpretation of the p value: A national survey study in academic psychologists from spain. *Psicothema*, *27*(3), 290–295. doi: 10.7334/psicothema2014.283

Bakker, M., Hartgerink, C. H. J., Wicherts, J. M., & van der Maas, H. L. J. (2016). Researchers intuitions about power in psychological research. *Psychological Science*. doi: 10.1177/0956797616647519

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*(6), 543–554. doi: 10.1177/1745691612459060

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews. Neuroscience*, *14*(5), 365–76. doi: 10.1038/nrn3475

Champely, S. (2009). Package "pwr." *October*, 1–21.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. United State of America: Lawrence Erlbaum Associates. Retrieved from http://www.utstat.toronto.edu/~brunner/oldclass/378f16/readings/CohenPower.pdf

Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, *45*(12), 1304-1312. doi: 10.1037/0003-066X.45.12.1304

Cohen, J. (1992a). A power primer. *Psychological Bulletin*, *112*(1), 155–159. doi: 10.1037/0033-2909.112.1.155

Cohen, J. (1992b). Statistical power analysis. *Psychological Science*. doi: 10.1111/1467-8721.ep10768783

Dufresne, R. J., Leonard, W. J., & Gerace, W. J. (2002). Making sense of students ' answers to multiple-choice questions. *The Physics Teacher*, *40*(March), 174–180.

Durlak, J. A. (2009). How to select, calculate, and interpret effect sizes. *Journal of Pediatric Psychology*, *34*(9), 917–928. doi: 10.1093/jpepsy/jsp004

Gigerenzer, G. (2004). Mindless statistics. *Journal of Socio-Economics*, *33*(5), 587–606. doi: 10.1016/j.socec.2004.09.033

Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E.-J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review*, *21*(5), 1157–1164. doi: 10.3758/s13423-013-0572-3

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124. doi: 10.1371/journal.pmed.0020124

Ioannidis, J. P. A. (2008). Why most discovered true associations are inflated. *Epidemiology*, *19*(5), 640–648. doi: 10.1097/EDE.0b013e31818131e7

Ioannidis, J. P. A., Munafò, M. R., Fusar-Poli, P., Nosek, B. A., & David, S. P. (2014). Publication and other reporting biases in cognitive sciences: Detection, prevalence, and prevention. *Trends in Cognitive Sciences*, *18*(5), 235–241. doi: 10.1016/j.tics.2014.02.010

Ioannidis, J. P. A., Ntzani, E. E., Trikalinos, T. A., & Contopoulos-Ioannidis, D. G. (2001). Replication validity of genetic association studies. *Nature Genetics*, *29*, 306–309. doi: 10.1038/ng749

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*(NOV), 1–12. doi: 10.3389/fpsyg.2013.00863

Lindsay, S. (2015). Replication in Psychological Science. *Psychological Science*, *26*(12), 1827–1832. doi: 10.1177/0956797615616374

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716-aac4716. doi: 10.1126/science.aac4716

Pluye, P., & Hong, Q. N. (2014). Combining the power of stories and the power of numbers: Mixed methods research and mixed studies reviews. *Annual Review of Public Health*, *35*(1), 29–45. doi: 10.1146/annurev-publhealth-032013-182440

R Core team. (2015). *A language and environment for statistical computing. R Foundation for Statistical Computing , Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-Project.org/.* Retrieved from http://www.mendeley.com/research/r-language-environment-statistical-computing-96/%5Cnpapers2://publication/uuid/A1207DAB-22D3-4A04-82FB-D4DD5AD57C28

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review*, *21*(2), 301–8. doi: 10.3758/s13423-014-0595-4

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*(11), 1359–1366. doi: 10.1177/0956797611417632

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, *143*(2), 534–547. doi: 10.1037/a0033242

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*(5), 779–804. doi: 10.3758/BF03194105

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p -values: Context, process, and purpose. *The American Statistician*, *70*(2), 129–133. doi: 10.1080/00031305.2016.1154108

Wickham, H. (2010). A layered grammar of graphics. *Journal of Computational and Graphical Statistics*, *19*(1), 3–28. doi: 10.1198/jcgs.2009.07098