A Methodological Critique of Focus on the "Average Student" in Psychological and Educational Research

Dr. Abdulrazaq A. Imam

Behavior Labs Africa Consultancy, and John Carroll University, USA, https://orcid.org/0000-0002-1262-6022

Abstract: Research in psychology and education tend to use large-N group designs that necessitate reporting of mean measures analyzed mostly with null hypothesis statistical testing (NHST), but sometimes with Bayesian, or the estimation approaches in inferential statistics. These approaches all render the average person or student as the the putative "subject" of interest in psychology and education, in addition to the myriad of serious problems, such as widespread replication failures, they have manifested. In reality, however, more often than not, it is the individual person or student who learns, not some nonexistent average person or student.A case is made that a largely ignored alternative to group designs is the Small-N or single-case designs that have a long and productive history in psychology, education, and medicine. They involve studying in-depth only a few subjects at a time under different conditions explored in some detail while observing systematic changes in behavior as those conditions change. In so doing, these designs not only focus on the individual, they reveal functional relationships between his or her behavior and the prevailing environmental conditions. In education, such environments range from the school, the classroom, or teacher (variables) to teaching methods, materials, and/or technology. Undoubtedly, adopting such designs more broadly in psychological and educational research would require a significant shift in how we approach asking questions, collecting data, analyzing and interpreting data, and making research and practice decisions. Not doing so, however, is like repeating the same mistake over and over and expecting a different result.

Keywords: Average measures, Average students, Methodology, Large-N group designs, Small-N single-subject designs

Citation: Imam, A. A. (2023). A Methodological Critique of Focus on the "Average Student" in Psychological and Educational Research. In M. Shelley, O. T. Ozturk, & M. L. Ciddi, Proceedings of ICEMST 2023--International Conference on Education in Mathematics, Science and Technology (pp. 330-344), Cappadocia, Turkiye. ISTES Organization.

Introduction

How often do we hear the phrase “average student” to categorize learners? Dutch students have a variety of answers for what the “average” means, some readily translatable in statistical terms, and others questionable (Bakker, 2003). Lay conceptions of what is average, therefore, may or may not be consistent with the technical

usage that we are all familiar with. When lay people use or hear the phrase "average student," it usually conjures up some comparative assessment (see Anuupadhyay, 2023), usually of academics. It often refers to learners as mediocre. If some are mediocre, of course, others are exceptional in whatever manner of intellectual attribute. Despite widespread use in daily life, the idea of an "average student" in psychology and education is largely based on statistical considerations rather than experimental or experiential necessity. So, such consolatory statements as "[a]verage people or students are also great. It's just that they may not get noticed and recognized as publicly" (Anuupadhyay, 2023) is sometimes warranted. We have to ask ourselves why there is so much focus on the "average person or student" in psychology and education if both disciplines presumably are essentially studies of the individual. The answer is rooted in the historical marriage of measurement and inferential statistics broadly speaking and particularly between that of measurement error and statistical control (i.e., as opposed to experimental control; see Cowles, 2001; Perone, 1991, 1999; Sidman, 1960).

History of Statistics in Psychology and Education

Academic achievement and performance are preeminently tied to measurement in both disciplines. The history of measuring human attributes in particular and nature more generally was characterized by challenges in discerning true measures and measurement errors due to the limitations imposed by our attempts to quantify and attach numbers to many attributes (Cowles, 2001; Stigler, 1992). According to Stigler, in regard to the entry of statistics into education, it was Francis Edgeworth's 1888 work that used the "normal distribution… as a scaling device" for exams (1992, p. 68). It is no wonder today that student performance is categorized based on the normal curve, readily having a place for the "average student" (see Anuupadhyay, 2023).

In an 1885 work on observations, Edgeworth was quoted as having stated: "Observations and statistics agree in being quantities grouped about a Mean; they differ, in that the Mean of observation is real, of statistics is fictitious" (Stigler, 1992, p. 67). At about the same time of Edgeworth's writings, Fechner and Charles Peirce were developing experimental attempts to quantify weight sensations in psychophysics. In replicating the earlier work of Fechner, Pierce had devised "a blind randomized experiment" with himself as subject and, in doing so, adopted "randomization to create an artificial baseline," which set the stage for "statistical investigation" in psychology ever since (Stigler, 1992, p. 65).

A most unfortunate aspect of the history of the linkage between statistical inference and experimental design is that most contemporary researchers believe they are operating a "unified theory of statistical inference" (Hubbard, 2004) in their use of null hypothesis statistical testing (NHST), when in fact it is a jumbled mishmash of technically incompatible statistical positions held by R. A. Fisher and Neyman-Pearson respectively (Schneider, 2015; see Imam, 2021, 2022). Imam (2022) described how psychology enjoys perhaps a unique status of having two research traditions, one using the ubiquitous snapshot group-designs approach to experimental investigations and the other relying on in-depth analyses of small numbers of subjects exposed to multiple conditions of experimental manipulation. The latter has been the saving grace for behavioral

psychology by escaping the embrace of inferential statistics' historical trap experienced by the former creating a deluge of averages in the literature (Imam, 2021).

As far back as the later decades of the 19th century, the average had been scrutinized. Writing contemporaneously with Edgeworth, Fechner, and Pierce mentioned earlier, the esteemed physiologist Claude Bernard (1927/1957) discussed the mistakes that can stem from physical, chemical, and biological averages. He argued that averages inevitably produce errors and therefore should be eschewed to avoid the muddling and misrepresentation they wield while "aiming to unify… and…to simply" (Bernard, 1927/157, p. 135).

Although the experimental psychologist may appear to need "statistical, as opposed to experimental control" (Cowles, 2001, p.20) due to perceived inability to achieve the latter on par with the physical sciences, "[a] major criticism of the effect of the use of statistical approach in psychological research is the failure to differentiate adequately between general propositions that apply to most, if not all, members of a particular group and statistical propositions that apply to some aggregated measure of the members of the group" (p. 19). Indeed, the apparent need for inferential statistics resulted from the emergence of "controlled experimentation in psychology" (Cowles, 2001, p. 172) in the 1920s and 1930s following R. A. Fisher's works on statistics and experimental design (see pp. 176-181).

Subject Matter and Subjects in Psychology and Education

Considering the definitions of psychology and education is a good place to begin a methodological critique of the phenomenon of the "average person or student" in psychology and education. The declared subject matters and the putative subjects in both disciplines are intricately linked. It is now standard to define psychology as the scientific study of behavior and mental processes, presumably, of the individual subject, human or animal (e.g., Bernstein et al., 2007). The focus on the individual is what differentiates psychology from sociology or anthropology. The definition of education is less precise and more varied (see Arslan , 2018), but a common theme is often that of learning and acquisition of knowledge (see Dragoescu, 2018). Again, by the individual subject, in this case though, only the human subject.

Despite the central interest in the individual person or student as the putative subject in psychology and education, the predominant experimental methodology in both disciplines is focused on averages. That is largely because of the heavy reliance on large-N group designs that are strictly tied to inferential statistics, which mostly take the form of null hypothesis statistical testing or NHST in both disciplines (e.g., Bohannon, 2014; Fidler et al., 2004; Gliner et al., 2002). The other two additional inferential approaches that are considerably less common in use are the estimation (e.g., Cumming & Calvin-Jageman, 2017) and the Bayesian (e.g., König & Schoot, 2018) approaches. At any rate, one thing the three approaches have in common is their reliance on the computation of the mean in order to make sense of data.

Statistics and Research Designs in Psychology and Education

The arithmetic mean (2008) is the most commonly used of the measures of central tendency, for good reason. When Babylonian astronomers of the third century BC first exploited it for planetary location, they could not have anticipated the impact or ubiquity of the arithmetic mean (2008) in the yet undefined social and psychological sciences of these days. And since then, in all the centuries before the entry of inferential statistics into modern research, the mean served prominently in the evaluation of data (see Smith et al., 2000;) and had not been smeared. Today, it is required for inference. The t-test and the F-test in the frequentist tradition, for example, are two common inferential statistics used in psychology and education to test hypotheses (see, e.g., Nestor & Schutt, 2015; Cohen et al., 2011). They estimate population parameters from sample statistics, usually the mean (e.g., Corty, 2016; Jackson, 2016; Runyon et al., 1996). For example, t = the difference in group means / total sample variance (or effect variance and error variance, respectively). As noted in the previous section, the use of inferential statistics is intricately linked to contemporary experimental design (see Jackson, 2016; Imam, 2021).

To illuminate on the pervasiveness of group designs and the attendant reliance on inferential statistics, let us consider the topical coverage of designs and statistics in research methods textbooks. A cursory survey of some such textbooks in psychology and education (see Table 1) reveals that, in both disciplines, research methods textbooks devoted: 1) substantially more space to group designs than to single-subject designs on the experimental design end, and 2) substantially more space to NHST than to Estimation and none at all to Bayesian approaches to inferential statistics on the quantitative methods end. A comparative assessment of the coverage in each discipline shows that qualitative methods in education received the most coverage of all followed by group designs in psychology, being the highest coverage in the discipline. Of the quantitative methods, there was negligible coverage of the estimation approach, but better than the zero coverage for the Bayesian approach in both disciplines. The table also shows that whereas there was more coverage of group designs and small-N designs in psychology compared to education, there was more coverage of qualitative methods, case studies , and NHST in education than in psychology in the textbooks.

Although NHST received the most coverage of the quantitative methods, all the three inferential statistical approaches require the mean for their respective roles in interpreting experimental data in psychology and education. Surprisingly, “only 17% (2 out of 12) of the texts” examined by Gliner et al. covered the doggedly widespread NHST controversies (2002, p. 89). This kind of findings probably account for why most researchers are oblivious to the ravages of the impact of NHST in psychological and educational research. Among other things, Sharpe posited “lack of awareness” as one factor responsible for researchers ignoring criticisms of NHST and other innovations recommended in their place (2013, pp. 573-574; see also Fidler et al., 2004). Not grasping the criticisms of NHST (see Lynch & Martin, 2017), researchers may be ignorant of the importance of the alternatives or assume they are all “equivalent” (Fidler et al., 2004, p. 120). The choice of textbooks may greatly inform what is taught or learned. Leaving out the controversies in textbooks, leaves readers ignorant and

uninformed about the costs and benefits of using the tools taught. In blissful ignorance, how do researchers conduct their work?

It is useful to consider the process of experimental research in psychology and education to better appreciate the need to be wary of the implications and side effects of the ubiquity of the mean measure in both disciplines. Typically, research begins with a formulation of a hypothesis (hopefully informed by theoretical considerations; see Szucs & Ioannidis, 2017; for contrary evidence), for which we select an appropriate experimental design (using experimental and control groups; Jackson, 2016), after which we recruit subjects (usually college students; Jaffe, 2005). Sometimes, a pilot study is conducted (ill-advisedly as a prelude to the real thing, see Sidman, 1960, pp. 217-233) at times badly (Francis, 2012), data collection then begins, and data analyzed (using NHST almost exclusively; see, e.g., Nickerson, 2000). The findings are then disseminated via conference presentation and/or publication. Finally, we hope, such findings contribute to theory building and more hypotheses (see, e.g., Oberauer & Lewadowsky, 2019).

It is equally important, however, to consider what is missing from how that process should work, in terms of basic methodological requirements. There is a problem of not satisfying the requirements of some fundamental statistical assumptions from the outset or of basic best practices. For example, specifying alpha and/or p value a priori as required from the outset (usually rarely done; see, e.g., Finch et al., 2001), specifying the relevant population of interest (required for the purpose of estimating parameters of the population; see Runyon et al.

Table 1. Number of pages (% of total/book and of psychology and education respectively) devoted to design and statistics topics in psychology and education research methods textbooks.

	Quantitative						
Book (edition)	NHST	Estimation	Bayesian	Qualitative	Group Designs	Case Study	Single-Subject
				Psychology			
Christensen et al., 2011 (11e)	35 (66%)	3 (6%)	0	15 (28%)	40 (63%)	3 (5%)	21 (33 %)
Cozby & Bates, 2012 (12e)	23 (77%)	1 (3%)	0	6 (20%)	59 (89%)	2 (3%)	5 (8%)
Nestor & Schutt, 2015 (2e)	12 (29%)	1 (2%)	0	28 (68%)	64 (71%)	0	26 (29 %)
Rosnow & Rosenthal, 2008 (6e)	83 (86%)	2 (2%)	0	11 (12%)	33 (89%)	0	4 (11 %)
Total	153 32%	7 1%	0	60 13%	196 41%	5 1%	56 12%

				Education			
Ary et al., 2010 (8e)	64 (40%)	4 (3%)	0	91 (57%)	59 (53%)	46 (41%)	7 (6%)
Cohen et al., 2011 (7e)	56 (43%)	3 (2%)	0	70 (54%)	25 (60%)	14 (33%)	3 (7%)
Fraenkel & Wallen, 2009 (7e)	54 (28%)	7 (4%)	0	132 (68%)	33 (53%)	2 (3%)	27 (44%)
Lodico et al., 2010 (2e)	16 (6%)	0	0	80 (83%)	27 (73%)	0	10 (27%)
Total	190 23%	14 2%	0	373 45%	144 17%	62 7%	47 6%

1996), or ensuring that population or data meet the normality requirement (rarely checked /usually unknown); see Bakker, 2014). Power analysis is often ignored (see, e.g., Finch et al., 2001) or inadequate (Button et al., 2013; Lodge et al., 2021; Stanczak et al., 2022; Weare, 2019). Random sampling is hardly ever done (relying instead on convenient samples of college students; see Jaffe, 2005; Grohol, 2010; Henrich et al., 2010).

Random assignment is sometimes flimsy (see Brown et al., 2023; Sella et al., 2021). So, just to reiterate, because of the pervasive use of group designs that require mean measures for experimental work in both disciplines, all the inferential statistical approaches treat the "average person or student" as the putative "subject" of interest (see Imam, 2022) in psychology and education. In reality though, more often than not, it is the individual person or student who learns, not some nonexistent "average person or student."

The use of the three statistical approaches, particularly NHST, has resulted in serious adverse consequences for psychology (see DeCoster et al ., 2015) and education. Such outcomes include rampant p-hacking (see Imam, 2018; Lindsay, 2015), replication failures (see Cesario, 2014; Makel & Plucker, 2014), lack of representativeness and generalizability (see Imam 2021; Jaffe, 2005), lack of a cumulative science (see Branch, 2014), a literature awash with massive psychological and educational averages (see Imam, 2022), to name a few. In the face of these dire consequences of the overwhelming reliance on large-N group designs, what else is left to do? The answer is in a largely ignored alternative to group designs that has a long and productive history in psychology, education, and medicine (see Bernard, 1927/1957; Moran & Malott, 2004; Sidman, 1960; Tankersley et al., 2008), namely, small-N single subject designs.

Basic Features of Small-N Designs in Psychology and Education

Small-N single-subject designs have served as the default method of investigating basic (Perone, 1991; Sidman, 1960) and applied (McLaughlin, 1983) processes in behavioral psychology since its inception in the early

decades of the last century (Iversen, 2013). Over the course of its development, behavioral psychology has become differentiated in its approach to the study of behavior, with the experimental analysis of behavior (EAB) focused on basic investigations of behavior and applied behavior analysis (ABA) focusing eminently on socially important behaviors. Both have thrived in their development as substantive areas in psychology to establish an independent research tradition that has avoided the trappings of inferential statistics experienced by the rest of psychology (Imam, 2021). To be clear, in the phrase small-N single subject design, if "the term 'single' describes the unit of analysis -the behavior of the individual- not the size of the sample," (Perone, 1991, p. 138), the small-N alludes to the sample size in contradistinction to the large-N requirement of group designs.

ABA has been the nexus for the introduction of small-N designs into the educational setting as a socially important setting (see McLaughlin, 1983; Sulzer-Azaroff & Mayer, 1994abc). Small-N single-subject designs focus solely on the individual, situation, or setting; mostly the individual, studying only a few of them at a time, each extensively exposed to various conditions of the relevant variables, each exposure lasting until measurement stability (see Perone, 1991). The interweaving of experimental and control conditions for each subject ensures that the individual experiences both, serving as his or her own control (McLaughlin, 1983). The real treat is that they reveal functional relationships between behavior and environmental conditions (McLaughlin, 1983; Perone, 1991; Sidman, 1960). In education, such environments can range from the school, the classroom, or teacher (variables) to teaching methods, materials, and/or technology (see Sulzer-Azaroff & Mayer, 1994abc). The following section provides three examples in the use of small-N single-subject designs in education to illustrate how small-N research works to preserve the ontological status of the individual in the research environment.

Illustrative Examples of Small-N Research in Education

The following examples show that small-N methodology is not all that foreign to educational research, as further attested to by the fact that some amount of space was devoted to their coverage in research methods textbooks in education (see Table 1). They represent only a sample of work that have been reported on educational topics in ABA.

In the first example, Witt and Elliott (1982) implemented a response cost lottery to manage student behavior in the classroom with minimal teacher resources. Three students previously exhibiting problem behaviors participated in an ABAB design. The results showed that for each child, appropriate on-task behaviors increased during each intervention (68% and 73%) relative to the respective baselines (of 10% and 43%).

Notably, the second baseline (showed more appropriate behavior (43%) than the first baseline (10%). The authors reported concomitant % changes in completed assignments: 27% for Baseline 1, 87% for Intervention 1, 38% for Baseline 2, and 90% for Intervention 2, supporting the effectiveness of the intervention in extending to other academically important behaviors.

In the second example, Munro and Stephenson (2009) reported on the use of active responding in a vocabulary classroom with 10-11 year olds of different nationalities. They compared hand raising (HR) to the use of response cards (RC) in different conditions in an ABAB (A = HR, B = RC) design. Across the conditions, they also recorded the teacher's questioning and feedback to students. The results showed that, on the part of the teacher, questioning was consistently at about the same rate throughout (about 1 response per minute or resp./min., under the first three conditions and 1.5 in the last) on the one hand, but feedback, on the other hand, increased when students used the response cards (to about 1.22 and 1.55 resp./min. under RC1 and RC2 respectively) relative to hand-raising (from .92 and .82 resp./min. under HR1 and HR2 respectively) baselines.

The students' results showed that active responding increased for each child when they used response cards (to about 85-100% and 90-100% under RC1 and RC2 respectively) compared to using hand raising (from about 10-30% and 12-28% respectively). Notably, the increase for Alice was not as high (to about 40-50% under both RC1 and RC2) as for the other students (perhaps because she was always at 0% during both of her baselines). If these kids had formed a response card experimental group in a group design, the group mean would have been reported, missing out on reporting the peculiarities of Alice's data in an average snapshot. Active responding has a positive effect on their test performances across the board, except for Alice (Nicky was absent for the final

test), again demonstrating extensions to other relevant academic behaviors. Their tests scores thus improved in the two corresponding HR-RC comparisons except for Alice's and Nicky's second comparisons; i.e., not counting Nicky's absence, indicating 88% improvement cases.

Finally, in the third example, Bohan and Smyth (2022) studied academically engaging and disruptive behaviors of two targeted students and the whole class of 9-10 year olds in an all-boys school in Ireland, using the Caught Being Good Game (CBGG) intervention. The design was an ABAB reversal design. For the whole class: 1) percentage of intervals with Academically Engaging Behavior (AEB) increased with CBGG relative to the baselines, and 2) Disruptive Behavior (DB) decreased with CBGG relative to the baselines. Notably. Only one data point overlapped during the first CBGG condition (with the initial baseline). The whole class as a unit of analysis thus exhibited orderly and consistent patterns of change as a function of the CBGG intervention.

At the individual level, for Adam, one of the two targeted students, 1) there was more variability in both baseline and under CBGG for both AEB and DB, compared to whole class behaviors, but 2) nevertheless, the results generally were consistent with the general functional effects of the contingencies, with generally higher percentage of AEB and lower DB under CBGG. For Ben, the second targeted student, 1) there was less variability in both baseline and under CBGG for both AEB and DB, and more consistent with the whole class, compared to Adam, and 2) DB was substantially higher compared to the whole class in baselines. The results of Bohan and Smyth demonstrate the value of individualizing data collection and analysis even when there is interest in the group as a whole.

Contrarian Approaches

To revisit the adverse outcomes emanating from the over reliance on group designs mentioned earlier, a comparison of each of the aforementioned outcomes under small-N design regimes shows that they are nonexistent in the small-N designs. Indeed, they represent contrarian approaches for psychological and educational research.

Table 2 provides the contrasting features of the two approaches in terms of their respective adverse impacts on the state of psychological and educational research practice. Whereas p-hacking is afforded by focus on statistical control in large-N designs, it is irrelevant and nonexistent in small-N designs that are focused on experimental control. The rampant failures of replication that has characterized NHST-based large-N designs are foreign to small-N designs because replications are built in by default both within and across conditions, as well as across subjects, settings, or situations. Because of the almost exclusive use of convenient samples, large N designs have been rendered lacking in representativeness and generalizability unlike in small-N designs in which the variety of replications ensure generality of reported effects. The problem of an evasive cumulative science under the large-N group design regimes becomes mute when small-N designs are implemented correctly. Finally, with a literature awash with averages as byproducts of the intersection of inferential statistics and experimentation in large-N group designs, which then implicates the putative subject of interest in psychology and education as the "average person or student," with small-N designs, the use of averages does not define the individual person or student.

Table 2. Contrarian approaches to experimental research in psychology and education Large-*N* Group

Designs	Small-*N* Single Subject Designs
p-Hacking (from focus on statistical control)	Utterly irrelevant (experimental control)
(rampant) Failures to replicate	Replications are built-in by default within and across conditions
• Lack of representativeness and generalizability functional (due to almost exclusive use of convenient samples)	Replication ensures generality (focus on relationships, not sampling
• Lack of a cumulative science	Mute issue
Literature awash with averages per person or student	Use of averages does not define the individual

As McLaughlin pointed out, small-N designs "avoid some of the obvious problems encountered in classroom action research such as control groups, randomization procedures, complex statistical analyses which are difficult for the classroom teacher to employ and carry out" (1983, p. 41). There then may be some incentives

for education researchers and practitioners to embrace small-N methodology more intentionally and reap the attendant benefits.

Implications of of Small-N Designs Acceptance

What are the implications of a wide acceptance of small-N single subject designs? First, it would focus attention on mastery in education, instead of perpetuating mediocrity by implication on account of the "average student." The ready acceptability of the notion of average student rooted in the ubiquity of the arithmetic mean in reporting research data is in part due to the extant nature of education, which educational systems tend to deliver in the classroom setting. That mode of delivery has meant that individual learners tend not to have one-on-one attention to their particular learning processes governed by their particular learning environment. Assessment of learning, therefore, have tended to be collective and comparative, and thus conveniently conducive to evaluation via the bell curve. Hence, the "average student."

Mastery is the hallmark of personalized systems of instruction (PSI), also known as "the Keller Plan" (Buterbaugh & Fuller, 1975; Fox, 2004). There is a long history in education with PSI, most commonly in hard sciences (e.g., Fuller, 1975), but now even more with instructional design advances afforded with computers (Fox, 20004) and artificial intelligence (AI). The typical PSI requires a proctoring tutor in addition to lectures designed to motivate the learner, among other things (Buterbaugh & Fuller, 1975; Fox, 2004). With AI, as Khan describes it, tutoring is not just for the student as it is in standard PSI, but help is provided for the teacher as well (2023, 0:49-0:58), bringing "the two sigma problem" to bear on raising the bar beyond "mastery" achieved with 1:30 ratio mastery learning to two standard deviations better performance with 1:1 ratio personal tutoring, transforming "average students into exceptional students and below average into average students" (see graphic, etc., Khan, 2023, 1:07-1:52) by using the Khan AI, Khanmigo. A whole different approach to education than the conventional one we are all familiar with, with a superior outcome, all centered on the individual learner.

Second, widespread adoption of small-N designs would require a significant shift in how we approach research in psychology and education. To begin with, asking questions would no longer require attending to pre-experimental statistical considerations such as a priori alpha and power (Finch et al., 2001) and parametric concerns about normality and homogeneity of variance (Bakker, 2014). The nature of data collection would change, from the snapshot approach aimed at computing group averages for comparison, to collecting data extensively on individual behavior for in-depth comparisons across conditions of manipulated variables (Perone, 1991). Analyzing and interpreting data would no longer be the exclusive purview of inferential statistics, which would retire to the remote circumstances where true group designs are warranted by the research questions, moving us closer to the behavioral reality of the individual instead. Finally, making research and practice decisions would be based on revealed functional relations, rather than statistical imperatives.

Conclusion

In conclusion, there is nothing that is inherently wrong with the average measure per see. Indeed, it has a long history that predates the invention of inferential statistics, some of which have come to symbolize its diminution as a quantitative tool. The real culprit is how we have been using it to make sense of data that we have encountered in our attempts to quantify nature and humanity. In good stead, we are not condemned to live with the "average student" side effect of our heavy reliance on large-N group designs, whose usually inadequate implementation has been responsible for the myriad of methodological problems including replication failures. There are demonstrably effective, viable alternatives in small-N single-subject designs that remove the cloak of mediocrity that inadvertently adorns many putative subjects in psychology and education. These designs ought to be more widely adopted for the myriad of methodological advantages they bring to the table. Not doing so is like repeating the same mistake over and over and expecting a different result.

Recommendations

- As researchers and educators, we should be wary of labeling learners as "average students."
- To do this most effectively, we need to be cognizant of the roots of its usage and how it is tied to statistics.

In deed, the idea is not based on experimental or experiential necessity.

- There are alternatives to NHST and large-N group designs for scientific research; consider small-N single subject designs for their myriad of methodological advantages for research and practice. The time and effort is well worth it.

References

Anuupadhyay (2023, April 12). Is it okay to be an average student? Retrieved from https://www.geeksforgeeks.org/is-it-okay-to-be-an-average-student/

Arithmetic Mean. (2008). In The concise encyclopedia of statistics. Springer. https://doi.org/10.1007/978-0-387- 32833-1_12

Arslan, H. (Ed.) (2018). An Introduction to Education. Cambridge Scholars Publishing.

Bakker, A. (2003). The early history of average values and implications for education. Journal of Statistical Education, 11, 1-18. https://doi.org/10.1080/10691898.2003.11910694

Bakker, M. (2014). Good science, bad science: Questioning research practices in psychological research. [Thesis]. Universiteit van Amsterdam.

Bernard, C. (1927/1957). An introduction to the study of experimental medicine. Dover Publications, Inc. Bernstein, D. A., Penner, L. A., Clarke-Stewart, A., & Roy, E. J. (2007). Psychology (7th ed.).

Houghton Mifflin.

Bohan, C., & Smyth, S. (2022). The Caught Being Good Game in a Middle Primary School Class. Behavior Analysis: Research and Practice, 22, 283–297. https://doi.org/10.1037/bar0000241

Bohannon, J. (2014). Replication effort provokes praise-and 'bullying' charges. Science, 344, 788-789. https://doi.org/10.1126/science.344.6186.788

Branch, M. (2014). Malignant side effects of null-hypothesis significance testing. Theory & Psychology, 24, 256-277. https://doi.org/10.1177/0959354314525282

Brown, M., McGrath, R. E., Bier, M. C., Johnson, K., & Berkowitz, M. W. (2023). A comprehensive meta-analysis of character education programs. Journal of Moral Education, 52, 119-138. https://doi.org/10.1080/03057240.2022.2060196

Buterbaugh, J. G., & Fuller, R. (1975). Personalized system of instruction (psi): An alternative. Personalized System of Instruction (PSI), or Keller Plan, Materials. 3. https://digitalcommons.unl.edu/physicspsikeller/3

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., et al. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. Nature Reviews Neuroscience, 14, 365-376. https://doi.org/10.1038/nrn3475

Cesario, J. (2014). Priming, replication, and the hardest science. Perspectives on Psychological Science, 9, 40-48. https://doi.org/10.1177/1745691613513470

Corty, E. R. (2016). Using and interpreting statistics: A practical text for the behavioral, social, and health sciences, 3e. Worth.

Christensen, L. B., Johnson, R. B., & Turner, L. A. (2011). Research methods, design, and analysis, 11e. Allyn & Bacon.

Cohen, L., Manion, L., & Morrison, K. (2011). Research methods in education, 7e. Routledge. Cowles, M. (2001). Statistics in psychology: An historical perspective. Lawrence Erlbaum.

Cozby, P. C., & Bates, S. C. (2015). Methods in behavioral research, 12e. McGraw Hill.

Cumming, G., & Calin-Jageman, R. (2017). Introduction to the new statistics: Estimation, open science, and beyond. Routledge.

DeCoster et al. (2015). Opportunistic biases their origins, effects, and an integrated solution. American Psychologists, 70, 499-514. https://doi.org/10.1037/a0039191

Dragoescu, A. A. (2018). Theories in language learning and teaching. In Arslan, H. (Ed.). (2018). An Introduction to Education, (pp. 11-22). Cambridge Scholars Publishing.

Fidler, F., Thomason, N., Cumming, G., Finch, S., & Leeman, J. (2004). Editors can lead researchers to confidence intervals, but can't make them think: Statistical reform lessons from medicine. Psychological Science, 15, 119-126. https://doi.org/10.1111/j.0963-7214.2004.01502008.x Finch, S., Cumming, G., & Thomason, N. (2001). Reporting statistical inference in the Journal of Applied Psychology: Little evidence of reform. Educational and Psychological Measurement, 61, 181-210. https://doi.org/10.1177/00131640121971167

Fox. E. J. (2004). The Personalized System of Instruction: A flexible and effective approach to mastery

learning. In D. J. Moran & R. W. Malott (Eds.), Evidence-based educational methods (pp. 201-221). Elsevier Academic Press.

Fraenkel, J. R., & Wallen, N. E. (2009). How to design and evaluate research in education, 7e. McGraw-Hill Francis, G. (2012). Publication bias and the failure of replication in experimental psychology. Psychonomic

Bulletin & Review, 19, 975-991. https://doi.org/10.3758/s13423-012-0322-y

Fuller, R. (1975). Introduction to the calculus-based physics modules. Personalized System of Instruction (PSI), or Keller Plan, Materials. 6. https://digitalcommons.unl.edu/physicspsikeller/6

Gliner et al. (2002). Problems with null hypothesis significance testing (NHST): What do the textbooks say?

Journal of Experimental Education, 7, 83-92. https://doi.org/10.1080/00220970209602058

Grohol, J. (2010). Psychology Secrets: Most Psychology Studies Are College Student Biased. Psych Central. Retrieved on August 15, 2017, from https://psychcentral.com/blog/archives/2010/08/26/psychology-secrets-mostpsychology-studies-are-college-student-biased/

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? Behavioral and Brain Sciences, 33, 61-135. https://doi.org/10.1017/S0140525X0999152X

Hubbard, R. (2004). Alphabet soup blurring the distinctions between p's and a's in psychological research.

Theory & Psychology, 14, 295–327. https://doi.org/10.1177/0959354304043638

Imam, A. A. (2018). Place of behavior analysis in the changing culture of replication and statistical reporting in psychological science. European Journal of Behavior Analysis, 19, 2- 10. https://doi.org/10.1080/15021149.2018.1463123

Imam A. A. (2021). Historically recontextualizing Sidman's Tactics: how behavior analysis avoided psychology's methodological Ouroboros. Journal of the Experimental Analysis of Behavior, 115, 115–28. https://doi.org/10.1002/jeab.661

Imam, A. A. (2022). Remarkably reproducible psychological (memory) phenomena in the classroom: some evidence for generality from small-N research. BMC Psychology, 10, 274-290. https://doi.org/10.1186/s40359-022-00982-7

Iversen, I. (2013). Single-case research methods: An overview. In G. J. Madden, W. V. Cube, T. D. Hackenberg, G. P. Hanley, & K. A. Lattal (Eds.), APA Handbook of behavior analysis: v. 1. Methods and principles (pp. 3-32). American Psychological Association. https://doi.org/10.1037/13937-000

Jackson, S. L. (2016). Research methods and statistics: A critical thinking approach. Cengage. Jaffe, E. (2005, September 10). How random is that? Observer, 18 (9). https://www.psychologicalscience.org/observer/how-random-is-that

Khan, S. (2023, April). How AI could save (not destroy) education [Video]. TED. https://youtu.be/hJP5GqnTrNo

König, C., & Schoot, R. (2018). Bayesian statistics in educational research: a look at the current state of affairs,

Educational Review, 70, 486-509, https://doi.org/10.1080/00131911.2017.1350636 Lindsay, D. S. (2015). Replication in psychological science. Psychological Science, 26, 1827-1932.https://doi.org/10.1177/0956797615616374

Lodge, J. M., Corrin, L., Hwang, G-J., & Thompson, K. (2021). Open Science and Educational Technology Research. Australasian Journal of Educational Technology, 37, 1-6. https://doi.org/10.14742/ajet.7565

Lodico, M. D., Spaulding, D. T., & Voegtle, K. H. (2010). Methods in educational research from theory to practice, 2e. Jossey-Bass.

Lynch, L., & Martin, N. (2017). Why 'what works' doesn't: False positives in education research. Retrieved from https://www.pearson.com/ped-blogs/blogs/2017/08/works-doesnt-false-positives-education-research.html

Makel, C. M., & Plucker, J. A. (2014). Facts are more important than novelty: Replication in the education sciences. Educational Researcher, 43, 304-316. https://doi.org/10.3102/0013189X14545513

McLaughlin, T. F. (1983). An examination and evaluation of single subject designs used in behavior analysis research in school settings. Educational Research Quarterly, 7, 35-42.

Moran, D. J., & Mallott, R. W. (Eds). (2004). Evidence-based Educational methods. Academic Press.

Munro, D. W., & Stephenson, J. (2009). The effects of response cards on student and teacher behavior during vocabulary instruction. Journal of Applied Behavior Analysis, 42, 795-800. https://doi.org/10.1901/jaba.2009.42-795

Nickerson, R. S. (2000). Null Hypothesis Significance Testing: A Review of an Old and Continuing Controversy. Psychological Methods, 5, 241-301. https://doi.org/10.1037/1082-989X.5.2.241

Nestor, P. G., & Schutt, R. K. (2015). Research methods in psychology: Investigating human behavior. 2e. Sage

Oberauer, K., & Lewandowsky, S. (2019). Addressing the theory crisis in psychology. Psychonomic Bulletin & Review, 26, 1596-1618. https://doi.org/10.3758/s13423-019-01645-2

Perone, M. (1991). Experimental design in the analysis of free-operant behavior. In I. Iversen & K. A. Lattal (Eds.), Experimental analysis of behavior, Part I (pp. 135-171). Elsevier.

Perone, M. (1999). Statistical inference in behavior analysis: Experimental control is better. The Behavior Analyst, 22, 190-116. https://doi.org/10.1007/BF03391988

Rosnow, R. L., & Rosenthal, R. (2008). Beginning behavioral research: A conceptual primer, 6e. Pearson Prentice Hall.

Runyon, R. P., Haber, A., Pittenger, D. J., & Coleman, K. A. (1996). Fundamentals of behavioral statistics. McGraw-Hill.

Schneider, J. W. (2015). Null hypothesis significance tests: A mix-up of two different theories—The basis for widespread confusion and numerous misinterpretations. Scientometrics, 102, 411-432. https://doi.org/10.1007/s11192-014-1251-5

Sella, F., Raz, G., & Kadosh, R. C. (2021). When randomisation is not good enough: Matching groups in intervention studies. Psychonomic Bulletin & Review, 28, 2085-2093. https://doi.org/10.3758/s13423-021-01970-5

Sharpe, D. (2013). Why the resistance to statistical innovations? Bridging the communication gap. Psychological Methods, 18, 572-582. https://doi.org/10.1037/a0034177

Sidman, M. (1960). Tactics of scientific research: Evaluating experimental data in psychology. Authors

Cooperative.

Smith, L. D., Best, L. A., Cylke, V. A., & Stubbs, D. A. (2000). Psychology without p values: Data analysis at the turn of the 19th century. American Psychologists, 55, 260-263. https://doi.org/10.1037/0003-066X.55.2.260

Stanczak, A., Darmon, C., Robert, A., et al. (2022). Do Jigsaw Classrooms Improve Learning Outcomes? Five Experiments and an Internal Meta-Analysis. Journal of Educational Psychology, 114, 1461-1476. https://doi.org/10.1037/edu0000730

Stigler, S. M. (1992). A historical view of statistical concepts in psychology and educational research. American Journal of Education, 101, 60-70. https://doi.org/10.1086/444032

Sulzer-Azaroff, B., & Mayer, G. R. (1994a). Achieving educational excellence: Behavior analysis for school personnel. Western Image.

Sulzer-Azaroff, B., & Mayer, G. R. (1994b). Achieving educational excellence: Behavior analysis for improving instruction. Western Image.

Sulzer-Azaroff, B., & Mayer, G. R. (1994c). Achieving educational excellence: Behavior analysis for achieving classroom and schoolwide behavior change. Western Image.

Szucs, D., & Ioannidis, J. P. A. (2017). When Null Hypothesis Significance Testing Is Unsuitable for Research: A Reassessment. Frontiers in Human Neuroscience, 11, 390. https://doi.org/10.3389/fnhum.2017.00390

Tankersley, M., Harjusola-Webb., & Landrum, T. J. (2008). Using single-subject research to establish evidence base of special education. Intervention in School and Clinic, 44, 83-90. https://doi.org/10.1177/1053451208321600

Weare, K. (2019). Mindfulness and contemplative approaches in education. Current Opinion in Psychology, 28, 321-326. https://doi.org/10.1016/j.copsyc.2019.06.001

Witt, J. C., & Elliott, S. N. (1982). The response cost lottery: A time efficient and effective classroom intervention. Journal of School Psychology, 20, 115-161. https://doi.org/10.1016/0022-4405(82)90009-7