Confidence vs Insight: Big and Rich Data in Computing Education Research

Neil C. C. Brown
King's College London
London, UK
neil.c.c.brown@kcl.ac.uk

Mark Guzdial
University of Michigan
Ann Arbor, MI, USA
mjguz@umich.edu

ABSTRACT

There are now many large datasets available for programming education research. They tend to be very large-scale, but often lack context or detailed participant information. This "big data" is in contrast to the "rich data" that has generally been collected from smaller, qualitative studies, with detailed context and participant information. Big data is often criticised for its lack of context, while rich data is often criticised for its small sample size which makes generalisable conclusions dubious. In this position paper we examine the constraints, advantages, and disadvantages of each type of data, and discuss how they can provide differing information on phenomena in programming education research. We argue that both types of data are useful and that we should value the potential findings of each, as well as encourage their combination in order to provide a complete picture of how people learn to program.

CCS CONCEPTS

• **Social and professional topics → Computing education**.

KEYWORDS

Big data, rich data

ACM Reference Format:
Neil C. C. Brown and Mark Guzdial. 2024. Confidence vs Insight: Big and Rich Data in Computing Education Research. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1 (SIGCSE 2024), March 20–23, 2024, Portland, OR, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3626252.3630813

1 INTRODUCTION

Programming education research is a significant part of computing education research [4, 42]. Research into programming education uses a wide variety of methods and datasets. There are qualitative methods used on small datasets, such as single-person case studies [11, 18, 31], detailed interview studies (e.g., the wonderfully titled "I like computers. I hate coding" [25]), observational studies [51], and survey studies that give us insights about student mindset [21, 32]. There are quantitative methods used on large datasets, such as examinations of compiler errors [8, 40], of student behaviour in learning environments [1, 15, 24], or bad programming habits [33].

In this position paper we are interested in a dichotomy of two different types of data: big data and rich data. Big data is large-scale (thousands to millions of participants), machine-collected and usually machine-analysed. Big data gives us *confidence* that the patterns observed generalise to a large population, but lacks human input and often any surrounding context about the participant. Rich data is human-collected, usually human-analysed, and full of detailed information that can give us *insight* into learners' intentions and the causal mechanisms. Both are now used in programming education research, but there has not been a detailed consideration of the advantages and disadvantages of big data, which is at risk of overshadowing rich data. Should we all be moving to big data? Or is big data a "busted flush" that cannot rival the value of rich data? In the rest of this paper, we discuss the advantages and disadvantages of big data and rich data. We provide an assessment of the limitations of each, and make recommendations to reviewers on when to discount studies and when not to, and recommendations to researchers on how we should be utilising the best of each.

2 RELATED WORK

Big data has been used in programming education research but there has not been much reflection in the literature on the impact of big data on programming education research. A simplistic search for the phrase "big data" in the ICER, ITiCSE, SIGCSE and TOCE proceedings (conducted in July 2023) turned up little in the way of reflective work about big data in research[1]: Hazzan and Shaffer [26] led a SIGCSE special session on big data in computer science education but there was not a full paper; Brown et al. [7] found that the Blackbox big dataset enabled some research that would not otherwise have been possible, but that many papers used a "somewhat shallow analysis technique."

Programming education researchers often analyse big data because programming data is so readily available. We can trace student behaviour at the level of keystrokes [45] and capture every version of a written program [8]. However, we may fall prey to the *McNamara fallacy*[2], where we believe those variables that we can easily measure are also the most important. Danielak [11] warned about the use of big data in his case study paper:

> "I argue that a growing syntax- and language-focused trend in computing education research is using bigger and richer datasets to push us farther away from the kind of fine-grained studies that helps understand design thinking in software creation."

[1] There were many results on how to *teach* about big data, which are not relevant here.
[2] https://en.wikipedia.org/wiki/McNamara_fallacy

SIGCSE 2024, March 20–23, 2024, Portland, OR, USA

ACM ISBN 979-8-4007-0423-9/24/03…$15.00
https://doi.org/10.1145/3626252.3630813

General education researchers also make choices between big and rich when potential data is available in machine-readable formats. They have explicitly explored the tensions, and have pointed out that big data is particularly valuable for studying process [2]. Fischer et al. [19] surveyed how researchers have used big data in general education research, concluding that big data is powerful but can miss fine-grained details. The *Journal of the Learning Sciences* has had several articles discussing the strengths and weaknesses of different data formats, such as video [13] and log data [20].

The difference between how the computing education research (CER) and learning sciences communities think about big and rich data can be seen in the handbooks for each field. The *Cambridge Handbook of Computing Education Research* includes a chapter on qualitative (rich) data analysis [50], but not on big data. The *Cambridge Handbook of the Learning Sciences* includes chapters on methods for analyzing big data (educational data mining [3]) as well as chapters on analyzing rich data (e.g., microgenetic analysis [44]).

There are whole conferences devoted to the analysis of big data, such as the *Educational Data Mining* (EDM) and the *Learning at Scale* (L@S) conferences. CER papers have appeared in the EDM conference [39], and there has even been a workshop on CS education research at EDM [56]. L@S has published studies on learning programming [52], often in the context of MOOCs with thousands of learners. At a panel at the 2014 L@S conference, Janet Kolodner schooled the community on the importance of learning sciences methods [41]. The more recent Learning at Scale conference proceedings have a mix of rich and big data analyses.

Our contribution is to explicitly ground discussion of these issues in computing education, with examples and challenges that are specific to programming education research. As we will describe in the next section, CER comes by big data more easily than most other education research subdisciplines, and inherits (from computer science) a more quantitative bent than general educational research, so it is a particularly interesting contrast in our discipline.

3 WHAT ARE THE CHARACTERISTICS OF BIG DATA AND RICH DATA?

In the following subsections we outline what we mean by big data and rich data in programming education research. We have deliberately picked colloquial names for these types of data, and have chosen to not get too hung up on exact definitions. The point of this paper is *not* to precisely and exhaustively classify every dataset into one or the other, but rather to discuss and contrast some commonly observed characteristics and features of these two kinds of dataset.

We deliberately focus on *data*, not on quantitative vs qualitative methods, or on philosophies of science. Of course, the analysis methods for big data tend to be quantitative and positivist, while the methods for rich data are much more varied. But our focus in this paper is on the characteristics of the data (we will explain later how this interacts with the choice of research questions, analysis methods and scientific approaches). For example, one of the typical limitations of big data is a lack of clear context of the participant, and no matter what research methodology you use – whether grounded theory or Hermeneutic phenomenology – you cannot recover what does not exist in the original data. So the data is the restriction, not the choice of subsequent analysis.

3.1 Big Data

Big datasets are, obviously, large. The scale usually involves thousands or even millions of participants. Big data is usually collected via a machine. The most common source of these datasets in CER are programming activity traces or sets of source code. Examples include the Scratch website [28], code.org activities [54] and Blackbox [8]. The datasets tend to be anonymous, with little to nothing known about the individual participants in terms of demographic- or meta-data (e.g., age, gender, age, background, educational status) which is often critical for CER [36]. Sometimes the demographics are known, but with little other information, such as a dataset of the exam results and demographics of all students in a country [29].

One of the reasons that computing education has been at the relative forefront of big data is that our discipline, especially programming, requires interaction with a computer. This makes it more technically feasible to observe people's activity. In contrast, imagine that we wanted to capture how mathematics students rearrange equations. This is commonly done on paper rather than on computer, so it would be much harder to collect a large dataset of this activity (unless it was turned into a machine-based activity [37]), compared to, say, refactoring in a programming environment.

3.2 Rich Data

Rich data refers to data which has been captured in a known context, usually involving contact with a human researcher, with useful personal information about each participant, and a lot of detail about each participant's activity. Rich data includes observation data (e.g., in a lab or classroom context [16]), interview data [25], and think-aloud protocols [10].

Rich data can be gathered from a large number of participants. Multi-institutional, multi-national (MIMN) studies in CER distribute data collection across many geographies and contexts [17], and can result in a significant dataset. Our definition still holds. In MIMN, there are humans interacting with other humans to gather data.

Rich data is typically analysed by humans, but this may change with future developments of technology. For example, computational grounded theory [34] involves computer-aided human analysis, which was used for example to analyse classroom audio recordings (rich data) [38] in a way that could potentially scale to large datasets in the future. This is another reason that we focus here on differences in data, rather than differences in analysis.

3.3 Can you have data that is big *and* rich?

Once we have defined big data and rich data, the obvious questions are whether they overlap and are exhaustive: can data be neither rich nor big – or both rich and big? Datasets can definitely be neither: a short evaluation survey with a few Likert scales collected after a 20-person workshop is neither big nor particularly rich. A survey can have thousands of people, with discrete responses (e.g., Likert scales) that are easily machine-read and analyzed, which meets our definition of big data. If a survey is administered in a known context (e.g., in a classroom) with open-ended responses, then it is rich, even though the survey might be anonymous.

In theory, data can be both big and rich, but in CER practice this is rarely the case. Rich data usually relies on a shared context or background, which is not true once you scale up: how often do

20,000 people share the same context beyond "they are users of tool X"? Rich data always requires (as part of our definition) human involvement in the collection and/or recording of the data. In theory human interviewers could run a study with hundreds or thousands of participants, but in practice this is very rare (see Seymour and Hewitt [43] for an example of interviewing 500 participants, which was then published as a 400+ page book) – it is hard to scale to big data sizes unless the data collection is completely automated.

3.4 Relation to other aspects of research

It is not strictly required that big data speaks to a positivist (or post-positivist) school of science, or that rich data is embraced by alternative schools, but in practice this is often true. To scale up to analyse millions of participants, analysis must involve automation rather than being solely conducted by humans on the raw data. Thus the analysis that can be performed on big data is limited to that which can be machine-encoded. Research questions which can be answered from an automated analysis tends towards positivist epistemologies and quantitative analysis. For example, it is hard to take a phenomenological (to study consciousness and human experience) stance if the data are a set of thousands of discrete responses to a survey. In a positivist framing, rich data are useful for understanding a context and defining future research questions for experiments which might be conducted with big data. This situation could change in future with developments in AI-aided human analysis.

The distinction between big and rich data is orthogonal to the issue of experimental design. Big data can be collected in experimental conditions, such as A-B testing in usability studies [30] or in non-experimental studies, such as observing all users of a programming tool [8]. Rich data can be gathered as part of an experiment, too. Many research designs, such as case studies or design-based research, will almost certainly involve collection of rich data because the number of participants is small and the context is knowable.

4 WHAT CAN WE KNOW FROM EACH?

Our ultimate aim in research is not to collect data, but to gain knowledge. The data is only important because of what we can learn from it. We posit that the knowledge we can get from each data type can be characterized as being about *confidence* or *insight*.

4.1 Big Data

The scale of big data has several distinct advantages for having *confidence* in our knowledge:

- Noise reduction: individual data points can always potentially be outliers, i.e., unusual examples that are unlike the average case. With a small data set there is a higher chance that one "noisy" outlier can bias the whole data set. Once the data is in the thousands or millions, this risk disappears.
- Comprehensiveness: a small sample from a large population has the risk that it may not contain certain rare behaviours. A larger-scale dataset is likely to pick up examples of all the different categories in the population.
- Reliability: a large sample is more likely to be representative of the entire population, giving more confidence that the results generalise to the whole population compared to a small study.
- Replicability: anyone could repeat the same (probably automated) analyses, despite not having been involved in the collection of the original data.

There are also other advantages that come from being machine collected:

- Unobtrusive recording: because big data is often logs of activity captured in a normal context (e.g., with commonly used tools in common settings) versus a laboratory setting, the data does not tend to suffer from demand characteristics (where participants are affected by the reactions and implied expectations of a human researcher). Instead, the data can be entirely uninfluenced by being in an experiment.
- Consistency: because big data is machine-collected (and usually machine-analysed), it is generally consistent in what is recorded (and how it is analysed). This is orthogonal from quality (it can be consistently poor or useless!) and it should also be recognised that it is not free from bias (see subsubsection 5.2.2).
- Anonymity: it can be easier to keep big data anonymous (simply by not collecting identifying information) and to convince participants that the data is anonymous, when they have no contact with a human researcher. However, just because it is machine collected does not guarantee anonymity – we will return to this in subsubsection 5.2.3.

All of these allow us, in (post-)positivist perspective, to make strong claims from analyses of big data. We can have some confidence that our claims about big data generalise to the overall population.

4.2 Rich Data

Rich data provides the opportunities to gain *insight* and has several advantages:

- Context: we can record details of the context in which data are collected when there is a human collecting the data. These might vary from a naturalistic description to a detailed description of the curriculum used in a classroom.
- Comprehensiveness: whereas big data is typically recorded from a tool and can miss other items such as classroom discussion or help-seeking, rich data can often record the complete activity and experience of participants, so that no details or aspects of their behaviour are missed.
- Identity: the possibility exists to gather critical demographic components of rich data that allow us to understand more about the identities of who is in our data set [36]. We might also gather data about students' prior experience and backgrounds, or how they are performing in the class.
- Nuance: subtleties of how context and identity interact can be present in rich data that cannot be in big data. Maybe an intervention works well for some students and not others (as in aptitude-treatment interactions [49]). Without rich data, we do not have those personal data with which we can correlate.
- Interpretation: rich data require a human to interpret them. A human can infer intentionality based on a theory of mind of the participant, and can posit causality about what context or identity variables led to a particular outcome. Without context or identity, big data offers few clues about causality or intentionality, so can often be difficult to interpret.

There is an important value of rich data with respect to the goals of diversity, equity, and inclusion that is not applicable to big data. If big data are collected without context and uniformly, the demographics of the participants are likely representative of the underlying population. If the population is biased (e.g., heavily skewed male, or towards wealthier students), then the big data will be too. If we want to understand something about a population that is minoritized or that is rarely present in the population, we will have to use human selection of the participants, which becomes rich (through the added context) and not big (because the deliberately-picked subset is unlikely to be large). An example is a recent paper by Wang et al. [53] where they gathered log file data in a summer learning experience with a small number of participants – that is rich data, but it is not big. If you are controlling participation, you are unlikely to get thousands or millions of participants.

If we want to study the experiences of a minoritized population, or to study what factors might help a part of the population to succeed, rich data are a better choice than big data to understand the nuances. Even if we had demographic information in a large dataset of programming activity, it does not seem complete to analyse by, say, gender, and find, for example, that male participants make different kinds of programming errors. We would need rich data to begin to understand the difference, or at least a combination – we will return to the idea of combination later.

5 WHAT ARE APPROPRIATE OR INAPPROPRIATE RESEARCH QUESTIONS?

Big data and rich data are each suited to different kinds of research questions, and have different pitfalls, which we will explore in the following subsections.

5.1 Big data

Big data is well-suited to answer questions such as "how often do people do *X*", "does *W* happen," or "what association is there between doing *X* and doing *Y*". The sheer amount of data can lead to reliable correlations on large-scale patterns in the data.

Big data is poorly-suited to answer questions such as "why do people do X", or "what do people think about X". Big data commonly captures only people's behaviour, but it is too much of a leap to infer their intention or opinions from behaviour alone.

5.2 Big data pitfalls

Big datasets are often limited in what information is available, either because it is impossible to collect it within the tool (e.g., when looking at a big dataset of shared online projects, it may be impossible to see the development process of each project) or because more information cannot be gathered *ad hoc* or *post hoc* – for example, a researcher may want to know why a participant took an unexpected action, but there is no way to ask the participant after the fact in a big data activity trace. This can lead to designing the research question around the available data, rather than the correct procedure of vice versa. Such research questions are often consistent and answerable, yet not very useful. They also often end up quite operational and low-level. For example, imagine that a researcher has access to a dataset of programming activity. A simple, answerable, but unimportant research question might be "how often do people write swear-words [curse words] in their editor?" By itself, this is mildly interesting but not very useful – except perhaps as a baseline for a better question. A more interesting question might be "how often do people swear in their code in response to an error, compared to during general coding?"

5.2.1 Correctness. One under-examined challenge of automated analysis on big data is ensuring correctness. With human analysis there is always the chance of individual mistakes (for example, mishearing a word in a transcript, or missing an important detail during a classroom observation) but systematic errors are less likely. However, with automated analysis there is the possibility of a bug in the analysis that produces entirely incorrect results. For example, imagine a source code analysis that inspected conditions of if-statements but accidentally overlooked all if-else statements (perhaps because they have a different node type in the abstract syntax tree). It would not be apparent to the researcher that this mistake had occurred, yet it could render all their results inaccurate.

5.2.2 Bias. There is a temptation to assume that because big data is collected and processed by machine, it must be free of bias. As mentioned earlier, it will be consistent – but it may well be consistently biased. Like all programs, the collection and analysis of big data reflects the biases of its creator. Studies have repeatedly found bias in systems based on big data [9, 14] and research using big data is similarly vulnerable to these issues. This may be in what is collected, how it is collected, or who it is collected from. If we collect no demographic information in big data then we cannot know if the dataset is biased. There are also issues around inferring patterns from big data [23] where inferring demographics (e.g., gender) from data can be biased and incorrect.

5.2.3 Anonymity. Big datasets can be hard to anonymise. With enough data, anyone can be identified [35]. One way to pursue anonymity is to simply not collect identifying data: if we do not ask for their name, age, gender, etc, then our data may become anonymous. But if there are enough clues in the program code, data may be de-anonymized. Activity logs of programming are ultimately free-text entry. There is the possibility that students have put their name in comments (likely for in-class assignments) but also in strings (e.g., to print their own name – or to test an address parsing function by using their real home address) or even in type names or variable names. This is not something that is easily solved automatically. In fact, we posit that recorded program code can only be considered anonymous if either:

- the language is so constrained that it does not allow user text entry of variables, names or strings (this may seem impossible, but environments like Light-bot [22] fall into this category), or
- the code has been checked (and redacted if necessary) by a human.

Fischer et al. [19] write "Finding the right balance between individual privacy and the public interest is very challenging... researchers face a choice between maximizing privacy and limiting the utility of the data set or maximizing utility but leaving the data subject to possible reidentification with sufficient effort." This applies to big data in all disciplines, including ours.

It is also important to consider that if the data is directly identifying (e.g., if their name is given) then under most ethical research

frameworks (and the principles of the GDPR in Europe), participants should have the right to withdraw their data. Managing this at a large scale can become complicated. Additionally, if a large data set contains personal data then the risks to participants of the dataset becoming public are increased.

5.3 Rich data

Rich data allow us to ask questions about individual identity, cognition, or affect, or about contextual factors. We can explore differences in process or outcomes based on the participant or the context. For example, we can ask if different genders or ethnicities engage in different behaviours, or if students with different backgrounds or abilities have different outcomes, preferences, or motivations. We can use rich data to ask questions like "under what conditions does X happen," "which students do Z," "do students with attribute Y have differences in behaviour or outcomes," or "is behaviour W more common with the experimental or control condition."

Rich data can be useful for understanding mechanism: "how does X happen?". This has a long history in medicine in the form of single-patient case studies, and it has also been used in computing education [12, 31].

Finally, rich data allow us to develop and test hypotheses about "why does X happen" or "under what conditions do we see outcome Y." With a small number of participants, we cannot be confident about the generalisability of the claims we make. A researcher can get to know one or a small number of participants and be able to suggest *why* the participants did what they did, or *how were they thinking or feeling* before they took particular actions. We can gain *insight* about causality and intentionality.

5.4 Rich data pitfalls

Rich data involves a relationship between human participants and human researchers. This relationship can end up influencing participants' responses. Especially in studies such as one-to-one interviews or case studies, the interviewer can (unintentionally) influence the responses they get from interviewees, and vice-versa. Participants can try to say what they think the researcher wants to hear, and researchers can form early opinions about participants that then bias how later data are gathered and analyzed [48].

Rich data can lead to becoming too enmeshed in the incredible level of detail available in the data: its richness can become overwhelming to the point where not enough attempt is made to simplify, abstract and generalise. This accords with the notion of Nuance Traps, described by Healy [27], such as the "fine-grain" trap, which is "a rejection of theory masquerading as increased accuracy." (ibid, page 120). If you become too focused on reporting the complete detail of the study and do not step back in order to summarise and transfer the knowledge, then you have provided only a detailed description and not a research finding.

It can be tempting with rich data to claim too much from a single interesting participant. An in-depth analysis of a single case can suggest causal mechanisms to be explored later, but it is unlikely to generalise. Generalisability should not be confused with saturation, which is a different objective for the sampling process [46]. The point of research is to develop knowledge, but that's not the same as theory. In rich data with a small number of participants, one can say absolutely that behaviours were observed and that participants offered explanations or rationale for those behaviours. The relationship between the rationale and the actions are concrete and observed. But are they predictive for others? Would other people apply the same rationale and make the same actions? Generalisation has different forms, like recognizing under what conditions a result might *transfer* [47]. Knowing the participant(s) and context well can lead to a nuance trap described as a "connoisseur". Healy describes them as follows [27, page 123]:

> Connoisseurs call for the contemplation of complexity almost for its own sake or remind everyone that things are subtler than they seem. The attractive thing about this move is that it is always available to the person who wants to make it. Theory is founded on abstraction, abstraction means throwing away detail for the sake of a bit of generality, and so things in the world are always "more complicated than that"—for any value of "that."

If the interaction with the participant was so rich that "you had to be there," it is not useful knowledge since not everyone was there. If the research can describe "there" in a way that others can see commonality or applicability of the observations, then the detailed description can be knowledge. It is important to simplify: to abstract away details that are not useful to understand the context. That makes the richness into knowledge that is useful to a research community.

The trick is to use rich data for what it's good for and big data for what it's good for. If we know a lot about the participants and the context, we can offer real insight – but it likely does not generalise in itself. Big data can give us great confidence, but can't tell us anything about context or individuals.

6 DISCUSSION

6.1 Weaving big and rich patterns

Our goal is science – understanding that allows us to construct explanations and make useful predictions. Science advances across multiple studies by multiple researchers. Understanding comes from insight, and larger datasets give us confidence about our predictions. We see value in weaving patterns of big and rich data across computing education research. One strand builds on the previous. All the strands together can form a tapestry (i.e., describe a scene) that no single strand can convey.

If we have a big data result that is interesting (e.g., this error is particularly common) but the reasons behind it are uncertain, then why not run a rich data study to find out? If we have a rich data result that is interesting but may not be representative of a larger population, why not use big data to find out? This seems obvious, yet we are not aware of many such interactions between the two worlds. We have only two examples of this, each from a different direction. In one example, a small set of interviews was analyzed and used to define a Likert-scale survey to be given to a large CS class, in order to determine how much the overall class was represented by the perspectives seen in the interviews [55]. In another example, big data was used to discover the most common mistakes from the activity of programming novices, which was

then used to survey educators (the known context making it rich data) on their predictions of the frequency of the mistakes [6].

One impediment is that some of the results that rich data finds may be difficult or impossible to operationalise as automated analysis. For example, rich data may find that students don't understand the definition of inheritance, but this is not obvious to turn into an automated analysis of some Java code. It may be that the other direction is more achieveable: to run a rich data study to investigate and deepen interesting findings from big data studies.

It is also valuable to synthesise existing results of big data and rich data studies on a particular topic, such as programmers' responses to compiler errors. This would necessarily be more of a narrative synthesis, since some review types such as meta-analyses have an inherent quantitative positivist restriction that would preclude inclusion of rich data papers which do not have numerical results to combine.

6.2 Imperfection and incompleteness can be acceptable

> "All models are wrong, but some are useful."
> – George Box [5, page 124]

Big data and rich data papers alike will always necessarily be incomplete in their view of the world, but we think that this should not preclude accepting them for publication[3]. We stress that we do not advocate for accepting bad research. There are good reasons to reject all kinds of papers, and bad research questions, inappropriate methodology or flawed execution are all valid reasons to reject a paper. However, the drive for high standards can be taken to the point of virtual impossibility, or to require several studies in each paper. We have discussed the idea of combining the best aspects of big data and rich data. But each of them, published alone, will always have flaws. Big data will always lack some context. Rich data will always have a small sample size. If we use these as reasons to reject then the only way to build on each is to publish big data and rich data together. This is would be an unproductive requirement: the bar for publication would be too high. Researchers would need to be skilled in both and, pragmatically, such a paper would not fit into the page limits. We must accept that all research will have limits to its knowledge, but not use this alone as a reason to prevent its publication. We should aim to be additive: publishing both quality big data and rich data research separately – rather than subtractive, excluding each solely because of their inherent limitations.

7 CONCLUSION

In this paper we have examined the role of big data and rich data in programming education research. We have given descriptions of each and explained their strengths and weaknesses: broadly, big data lacks context while rich data lacks scale. Each type has pitfalls to fall into, with big data risking shallow or constricted analysis, and rich data risking over-interpretation and lack of generalisation.

We have characterized the tension between big and rich as the difference between *confidence* and *insight*, but going beyond pithy one-word summaries, we considered *what* we can have confidence or insight about. We can only be absolutely confident about observed behaviour over time. Big data can tell us with great confidence that actions were taken and their sequentiality. Cognition and emotion are invisible and impossible to measure directly, but only rich data can get us close. We can never be confident in claims about what students thought or felt. Our explanations about causality and mechanism are based on student knowledge, intention, motivation, and emotion, which provides *insight*.

We make the argument here that despite inherent limitations, both big data and rich data are useful, especially if their limitations are understood both by researchers and reviewers. We have offered pathways for how to combine the two: not by turning all individual studies into hybrid studies, but rather by allowing studies from both sides of the "divide" to inform each other in hypothesis generation and explanatory power, and/or using synthesis papers to combine the results of big data and rich data papers.

7.1 Recommendations for reviewers

Our overall message is one of respect for each other's work, and recognition of inherent constraints. Some constraints are impossible to overcome (interview studies will never have thousands of participants; huge datasets will never involve individual human contact with participants) but that alone is not a reason to reject. Of course, we do not recommend accepting all papers. Within any study there can be fatal flaws in the design or execution.

We offer the following recommendations for reviewers:

- We should not reject big data papers because the data lacks context, if the context is not important for the research question.
- We should not reject rich data papers for having a small sample size. Even a single-participant case study can have value if the analysis tells us something new and valuable (e.g., studies by Danielak [11], and Fincher and Tenenberg [18]).
- We should not require completeness in a single study. There are cases where neither big nor rich data can provide a full picture alone, but it is unreasonable to require both in a single paper.
- We should reject papers that ask research questions that cannot be answered by their data.
- We should reject papers that use a methodology that is inappropriate for their data, such as choosing statistical analysis on rich data without sufficient statistical power or making claims about context or identity with big data.

ACKNOWLEDGMENTS

We are grateful to Brian Danielak and the anonymous reviewers.

REFERENCES

[1] Efthimia Aivaloglou and Felienne Hermans. 2016. How Kids Code and How We Know: An Exploratory Study on the Scratch Repository. In *ICER 2016*. ACM, 53–61. https://doi.org/10.1145/2960310.2960325
[2] Maria Ijaz Baig, Liyana Shuib, and Elaheh Yadegaridehkordi. 2020. Big data in education: a state of the art, limitations, and future research directions. *International Journal of Educational Technology in Higher Education* 17, 1 (2020), 1–23.
[3] Ryan S Baker, Taylor Martin, and Lisa M Rossi. 2016. Educational data mining and learning analytics. *The Wiley handbook of cognition and assessment: Frameworks, methodologies, and applications* (2016), 379–396.
[4] Brett A. Becker and Keith Quille. 2019. 50 Years of CS1 at SIGCSE: A Review of the Evolution of Introductory Programming Education Research. In *SIGCSE 2019*. ACM, 338–344. https://doi.org/10.1145/3287324.3287432
[5] George EP Box and Norman R Draper. 1987. *Empirical model-building and response surfaces*. John Wiley & Sons.

[3] Anecdotally, we believe that both kinds of paper get rejected for being incomplete, although it is hard to assess the scale of the problem because reviews are always private and we do not see rejected papers.

[6] Neil C. C. Brown and Amjad Altadmri. 2017. Novice Java Programming Mistakes: Large-Scale Data vs. Educator Beliefs. *ACM Trans. Comput. Educ.* 17, 2, Article 7 (may 2017), 21 pages. https://doi.org/10.1145/2994154

[7] Neil C. C. Brown, Amjad Altadmri, Sue Sentance, and Michael Kölling. 2018. Blackbox, Five Years On: An Evaluation of a Large-Scale Programming Data Collection Project. In *ICER 2018*. ACM, 196–204. https://doi.org/10.1145/3230977.3230991

[8] Neil C. C. Brown, Michael Kölling, Davin McCall, and Ian Utting. 2014. Blackbox: A Large Scale Repository of Novice Programmers' Activity. In *SIGCSE 2014*. ACM, 223–228. https://doi.org/10.1145/2538862.2538924

[9] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. PMLR, 77–91.

[10] Kathryn Cunningham, Rahul Agrawal Bejarano, Mark Guzdial, and Barbara Ericson. 2020. "I'm Not a Computer": How Identity Informs Value and Expectancy During a Programming Activity. (2020). https://doi.org/10.22318/icls2020.705

[11] Brian Danielak. 2022. How Code Takes Shape: Studying a Student's Program Evolution. *Cognition and Instruction* 40, 2 (2022), 266–303.

[12] Brian A Danielak, Ayush Gupta, and Andrew Elby. 2014. Marginalized identities of sense-makers: Reframing engineering student retention. *Journal of Engineering Education* 103, 1 (2014), 8–44.

[13] Sharon J. Derry, Roy D. Pea, Brigid Barron, Randi A. Engle, Frederick Erickson, Ricki Goldman, Rogers Hall, Timothy Koschmann, Jay L. Lemke, Miriam Gamoran Sherin, and Bruce L. Sherin. 2010. Conducting Video Research in the Learning Sciences: Guidance on Selection, Analysis, Technology, and Ethics. *Journal of the Learning Sciences* 19, 1 (2010), 3–53. https://doi.org/10.1080/10508400903452884

[14] Catherine D'ignazio and Lauren F Klein. 2023. *Data feminism*. MIT press.

[15] Deborah A. Fields, Yasmin B. Kafai, and Michael T. Giang. 2017. Youth Computational Participation in the Wild: Understanding Experience and Equity in Participating and Programming in the Online Scratch Community. *ACM Trans. Comput. Educ.* 17, 3, Article 15 (Aug. 2017), 22 pages. https://doi.org/10.1145/3123815

[16] Sally Fincher, Sebastian Dziallas, and Daniel Knox. 2019. Space, Place and Practice in Computing Education. In *UKICER 2019*. ACM, Article 11, 7 pages. https://doi.org/10.1145/3351287.3351297

[17] Sally Fincher, Raymond Lister, Tony Clear, Anthony Robins, Josh Tenenberg, and Marian Petre. 2005. Multi-Institutional, Multi-National Studies in CSEd Research: Some Design Considerations and Trade-Offs. In *ICER 2005*. ACM, 111–121. https://doi.org/10.1145/1089786.1089797

[18] Sally Fincher and Josh Tenenberg. 2007. Warren's Question. In *ICER 2007*. ACM, 51–60. https://doi.org/10.1145/1288580.1288588

[19] Christian Fischer, Zachary A. Pardos, Ryan Shaun Baker, Joseph Jay Williams, Padhraic Smyth, Renzhe Yu, Stefan Slater, Rachel Baker, and Mark Warschauer. 2020. Mining Big Data in Education: Affordances and Challenges. *Review of Research in Education* 44, 1 (2020), 130–160. https://doi.org/10.3102/0091732X20903304

[20] Janice D Gobert, Michael Sao Pedro, Juelaila Raziuddin, and Ryan S Baker. 2013. From log files to assessment metrics: Measuring students' science inquiry skills using educational data mining. *Journal of the Learning Sciences* 22, 4 (2013), 521–563.

[21] Jamie Gorson and Eleanor O'Rourke. 2020. Why do CS1 Students Think They're Bad at Programming? Investigating Self-efficacy and Self-assessments at Three Universities. In *ICER 2020*. 170–181.

[22] Lindsey Ann Gouws, Karen Bradshaw, and Peter Wentworth. 2013. Computational Thinking in Educational Activities: An Evaluation of the Educational Game Light-Bot. In *ITiCSE 2013*. ACM, 10–15. https://doi.org/10.1145/2462476.2466518

[23] Catherine Greene. 2019. Big Data and the Reference Class Problem. What Can We Legitimately Infer about Individuals? *Computer Ethics-Philosophical enquiry (CEPE) proceedings* 2019, 1 (2019), 7.

[24] Philip J. Guo. 2018. Non-Native English Speakers Learning Computer Programming: Barriers, Desires, and Design Opportunities. In *CHI 2018*. ACM, 1–14. https://doi.org/10.1145/3173574.3173970

[25] Paulina Haduong. 2019. "I like computers. I hate coding": a portrait of two teens'experiences. *Information and Learning Sciences* 120, 5/6 (2023/03/23 2019), 349–365. https://doi.org/10.1108/ILS-05-2018-0037

[26] Orit Hazzan and Clifford A. Shaffer. 2015. Big Data in Computer Science Education Research. In *SIGCSE 2015*. ACM, 591–592. https://doi.org/10.1145/2676723.2677318

[27] Kieran Healy. 2017. Fuck nuance. *Sociological Theory* 35, 2 (2017), 118–127.

[28] Benjamin Mako Hill and Andrés Monroy-Hernández. 2017. A longitudinal dataset of five years of public activity in the Scratch online community. *Scientific Data* 4, 1 (31 Jan 2017), 170002. https://doi.org/10.1038/sdata.2017.2

[29] Peter E. J. Kemp, Billy Wong, and Miles G. Berry. 2019. Female Performance and Participation in Computer Science: A National Picture. *ACM Trans. Comput. Educ.* 20, 1, Article 4 (Nov. 2019), 28 pages. https://doi.org/10.1145/3366016

[30] Rochelle King, Elizabeth F Churchill, and Caitlin Tan. 2017. *Designing with data: Improving the user experience with A/B testing*. " O'Reilly Media, Inc.".

[31] Colleen M. Lewis. 2012. The Importance of Students' Attention to Program State: A Case Study of Debugging Behavior. In *ICER 2012*. ACM, 127–134. https://doi.org/10.1145/2361276.2361301

[32] Alex Lishinski, Aman Yadav, Jon Good, and Richard Enbody. 2016. Learning to program: Gender differences and interactive effects of students' motivation, goals, and self-efficacy on performance. In *ICER 2016*. ACM, 211–220.

[33] Jesús Moreno and Gregorio Robles. 2014. Automatic detection of bad programming habits in scratch: A preliminary study. In *FIE 2014*. IEEE, 1–4.

[34] Laura K. Nelson. 2020. Computational Grounded Theory: A Methodological Framework. *Sociological Methods & Research* 49, 1 (2020), 3–42. https://doi.org/10.1177/0049124117729703

[35] Paul Ohm. 2009. Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA l. Rev.* 57 (2009), 1701.

[36] Alannah Oleson, Benjamin Xie, Jean Salac, Jayne Everson, F Megumi Kivuva, and Amy J Ko. 2022. A Decade of Demographics in Computing Education Research: A Critical Review of Trends in Collection, Reporting, and Use. In *ICER 2022*. 323–343.

[37] Erin Ottmar, David Landy, Erik Weitnauer, and Rob Goldstone. 2015. Graspable mathematics: Using perceptual learning technology to discover algebraic notation. In *Integrating touch-enabled and mobile devices into contemporary mathematics education*. IGI Global, 24–48.

[38] Chris Palaguachi, E Cox, and C D'Angelo. 2022. Audio Analysis of Teacher Interactions with Small Groups in Classrooms. In *General Proceedings of the 15th International Conference on Computer-Supported Collaborative Learning 2022*.

[39] Thomas Price, Rui Zhi, and Tiffany Barnes. 2017. Evaluation of a Data-Driven Feedback Algorithm for Open-Ended Programming. *International Educational Data Mining Society* (2017).

[40] David Pritchard. 2015. Frequency Distribution of Error Messages. In *PLATEAU 2015*. ACM, 1–8. https://doi.org/10.1145/2846680.2846681

[41] Mehran Sahami, Jace Kohlmeier, Peter Norvig, Andreas Paepcke, and Amin Saberi. 2014. Panel: Online Learning Platforms and Data Science. In *L@S 2014*. ACM, 137–138. https://doi.org/10.1145/2556325.2579110

[42] Carsten Schulte. 2013. Reflections on the Role of Programming in Primary and Secondary Computing Education. In *WiPSCE 2013*. ACM, 17–24. https://doi.org/10.1145/2532748.2532754

[43] Elaine Seymour and Nancy M Hewitt. 1997. *Talking about leaving: why undergraduates leave the sciences*. Westview Press, Boulder, CO.

[44] Bruce L. Sherin and Clark A. Chinn. 2022. *Microgenetic Methods* (3 ed.). Cambridge University Press, 217–237. https://doi.org/10.1017/9781108888295.014

[45] RajShrestha Shrestha, Juho Leinonen, Albina Zavgorodniaia, Arto Hellas, and John Edwards. 2022. Pausing While Programming: Insights from Keystroke Analysis. In *ICSE-SEET 2022*. Association for Computing Machinery, New York, NY, USA, 187–198. https://doi.org/10.1145/3510456.3514146

[46] Mario Luis Small. 2009. 'How many cases do I need?': On science and the logic of case selection in field-based research. *Ethnography* 10, 1 (2009), 5–38. https://doi.org/10.1177/1466138108099586

[47] Brett Smith. 2018. Generalizability in qualitative research: misunderstandings, opportunities and recommendations for the sport and exercise sciences. *Qualitative Research in Sport, Exercise and Health* 10, 1 (2018), 137–149. https://doi.org/10.1080/2159676X.2017.1393221 arXiv:https://doi.org/10.1080/2159676X.2017.1393221

[48] Joanna Smith and Helen Noble. 2014. Bias in research. *Evidence-based nursing* 17, 4 (2014), 100–101.

[49] Richard E Snow. 1991. Aptitude-treatment interaction as a framework for research on individual differences in psychotherapy. *Journal of consulting and clinical psychology* 59, 2 (1991), 205.

[50] Josh Tenenberg. 2019. *Qualitative Methods for Computing Education*. Cambridge University Press, 173–207. https://doi.org/10.1017/9781108654555.008

[51] Sara Vogel. 2020. *Translanguaging About, With, and Through Code and Computing: Emergent Bi/Multilingual Middle Schoolers Forging Computational Literacies*. Ph. D. Dissertation.

[52] Qianxiang Wang, Wenxin Li, and Tao Xie. 2014. Educational Programming Systems for Learning at Scale. In *L@S 2014*. ACM, 177–178. https://doi.org/10.1145/2556325.2567868

[53] Wengran Wang, Yudong Rao, Archit Kwatra, Alexandra Milliken, Yihuan Dong, Neeloy Gomes, Sarah Martin, Veronica Catété, Amy Isvik, Tiffany Barnes, Chris Martens, and Thomas Price. 2023. A Case Study on When and How Novices Use Code Examples in Open-Ended Programming. In *ITiCSE 2023*. ACM, 82–88. https://doi.org/10.1145/3587102.3588774

[54] David Weintrop, Heather Killen, Talal Munzar, and Baker Franke. 2019. Block-Based Comprehension: Exploring and Explaining Student Outcomes from a Read-Only Block-Based Exam. In *SIGCSE 2019*. ACM, 1218–1224. https://doi.org/10.1145/3287324.3287348

[55] Svetlana Yarosh and Mark Guzdial. 2007. Narrating data structures: the role of context in CS2. In *ICER 2007*. ACM, New York, NY, USA, 87–98. https://doi.org/10.1145/1288580.1288592

[56] Rui Zhi, Thomas W Price, Nicholas Lytle, Yihuan Dong, and Tiffany Barnes. 2018. Reducing the state space of programming problems through data-driven feature detection. In *Educational Data Mining in Computer Science Education (CSEDM) Workshop at EDM*.