Research Evaluation, volume 13, number 3, December 2004, pages 189–195, Beech Tree Publishing, 10 Watford Close, Guildford, Surrey GU1 2EP, England

Experimental repetition

What is the value of replicating other studies?

Caroline L Park

In response to a question on the value of replication in social science research, the author undertook a search of the literature for expert advise on the value of such an activity. Using the information gleaned and the personal experience of attempting to replicate the research of a colleague, the conclusion was drawn that replication has great value but little 'real life' application in the true sense. The activity itself, regardless of the degree of precision of the replication, can have great merit in extending understanding about a method or a concept.

Dr Caroline L Park, RN, is Associate Professor, Centre for Nursing and Health Studies, Athabasca University, 1 University Drive, Athabasca, Alberta T6S 3A3, Canada; tel: 1-800-788-9041 ext. 6381; email: clpark@athabascau.ca

LAST YEAR I NAÏVELY SUBMITTED a proposal for funding review indicating that I was going to replicate the study of a colleague. When the review comments came back to me one of them was: 'What is the value of replication?' I was initially quite offended by this comment. Not having any nonverbal clues with which to associate the question, I assumed it was facetious. The answer is self-evident, isn't it?

I have learned, in my later years, to mull things over. A few days later I started wondering if I had interpreted the comment correctly. This wondering led me to the literature, starting with Neuliep's (1990) *Handbook of Replication Research in the Behavioural Social Sciences*. Now that I've read the views of many experts, and have replicated my colleague's research, I feel better prepared to answer what I now consider was a serious question.

The lone researcher

The basic reason that research must be replicated is because the findings of a lone researcher might not be correct. Rosenthal (1990) wrote that unreplicated research might reflect unidentified error caused by: undetected equipment failure; rare, possibly random human errors of procedures — observation, recording, computations or reporting; the results might be a random 'fluke' and/or the outcomes might reflect individual organismic differences and/or systemic experimental effects. Or, as Miller had said in 1980 (cited by Lamal in 1990), 'The fact that a theory has passed one test provides no evidence at all that it will pass a repetition of the test.' We have reached a point in social science research where some feel that 'the literature is replete with one-shot studies of phenomenon whose veracity is unquestioned and

 0958-2029/04/030189-7 US$08.00

whose findings are disseminated as implicit laws' (Easley *et al*, 2002, page 83). This strong assumption is not echoed by all. Many qualitative researchers feel that there have been enough studies in their concept field to support practice and yet, their theory is still questioned and the findings are not disseminated.

Research that is totally quantitative — i.e. weights and measures — can be replicated with great accuracy or precision. This is the basic assumption behind high school chemistry labs. Under optimum conditions, with specific directions and pure ingredients, most high school students succeed in replicating the research assignment. So, we know it is possible. The theory being tested is supported again and again and becomes part of our general knowledge system. Regardless, even these 'quantitative' experiments are not always successfully replicated, for a myriad of reasons, some of which have been cited already. Some of us were never able to get it 'right' in the chemistry lab but the research *was* replicable.

Precision, or accuracy, of a replication to imitate the research under question is much more difficult to obtain when we move away from weights and measures, and into behavioural science. The more qualitative a study, the more difficult it is to show replicability, let alone to replicate. Does that mean that it is of no value to try?

Precision of replication

According to Rosenthal (1990) replications are 'relative', depending upon how close they are to the original study, in terms of subjects, experimenters, tasks and situations. One purpose of replication, or an attempt to replicate, is to strengthen the foundation of a theory so that it too will become a part of our general knowledge system. Another is to test its veracity or truth to determine if it should be supported as knowledge at all. Table 1 shows the effects on this knowledge-building of successful and unsuccessful replications. It does not matter if the replication is precise or not. It can still have an effect and this effect can be positive or negative. This means that both precise and imprecise replications have value.

Table 1. Precision vs. success

	Successful replication	Unsuccessful replication	Effect of unsuccessful replication on original investigator
Precise replication	Supports the theory	Damages the theory	Impugns the investigator
Imprecise replication	Extends the theory	Limits the theory	Impugns the investigator very little

Table 1 introduces an interesting concept: that of damage to a researcher's reputation if their work is not replicable. Successful replication and building of a theory is obviously advantageous to a researcher's reputation. Neither of these will happen if no one ever tries to replicate another researcher's work.

> When a replication successfully confirms the findings of the original study it proves at least some support for the theory concerned. When a theory repeatedly fails to be replicated it is more plausible to regard the original findings as a result of chance factors or idiosyncrasies of the context, rather than the manifestation of real structures or mechanisms. (Tsang and Kwan, 1999, page 770)

Amir (1990) reiterated the assumption, which I believe is widely held, that before a result can serve as a basis for theory it must pass tests of reproductability and generalizability. 'Reproductability' means that a particular psychological aspect or phenomenon found to occur in a certain sample occurs in similar samples, and 'generalizability' means whether it occurs in different types of groups as well. Reproductability is replicability. Generalizability goes a step further but also supports the theory and is replication.

'Implicit in all our discussion of replication is the idea that the original study is worth replicating' (Rosenthal, 1990, page 7). Rosenbaum (2001) indicates that poorly designed studies tend to replicate even when incorrect. He also cautions against replicating a hidden bias.

Determining the precision of a replication

Hendrick (1990) provided eight aspects to consider in replication research (see Table 2). I thought about each of these in relation to my own replication of Garrison *et al*'s (2000) 'cognitive presence' research. The original study is a content analysis of students' postings to discussion boards in online university courses to assess if a coding tool 'is a practical approach to judge the nature and quality of discourse in a computer conference' (page 4). The postings are coded using a hierarchical listing of cognitive levels. The study appeared to be quantitative, in that content was counted and calculated into comparison figures. In fact, the definitions of the categories of cognition are very subjective and the coding of student postings into a specific category is likewise.

Hendrick (1990) also discusses four types of replication: conceptual, partial, exact and systemic. Using the data in Table 2, I would have to assume that my replication was partial because there are distinct differences in some of the aspects but also some marked similarities.

Table 2. Hendrick's aspects: similarities of two studies

Aspect	Original study	My replication
1. subject characteristics	Graduate students in two different courses at a traditional university participated in a discussion conference as part of a course. There is no data on demographics of students, or prior learning experiences	Graduate students in one totally online course in a distance virtual university conducted all of their communication, student to student and student to faculty on line. There is no data on demographics of students, or prior learning experiences
2. specific research histories of subjects – prior experiences and or methods used to get them into the study	– previous research history not known – volunteers solicited via email. Self-select from within a group	– senior course, students had experience with technology and online conferencing – consent form built into courses – all students in the course participated
3. historical context – social cultural systemic factors relating to time and place	– web-based course within larger, urban, bricks-and-mortar university. The participants had the opportunity to meet and talk outside of the discussion	– totally online course at distance university. The students never met face-to-face
4. general physical setting of the research 5. control agent 6. specific task variables	– online conferencing – none – web C. T. course – two faculty – different instructions	– online conferencing – none – web C. T. course – one faculty – different instructions
7. primary information focus	– cognitive categories developed by team – definition of terms by team	– cognitive categories developed by original researcher – definition of terms adjusted to reflect the understanding of this team
8. modes of data reduction in presentation	– Atlas.ti for content analysis – coders selected and trained by this team – potentially different interpretations of the codes	– Atlas.ti for content analysis – coders selected and trained by this team – potentially different interpretations of the codes

Experimenter effect

Another important concept raised by Rosenthal is 'experimenter effect'. He believes that not only are experimenters with different research interests different kinds of people, and that different kinds of people are likely to obtain different data from their subjects, but also that researchers can hold an area of interest in common but hold opposite expectancies about the results of any given experiment. When it comes to team research, he believes it is reasonable to assume that colleagues or faculty and students in the same department will be highly correlated by both natural selection and training factors. Morse and Mitcham (2002) describe a similar observation more recently, in the language of qualitative research as 'conceptual tunnel vision'. Researchers often over-categorize their data, assigning more data to the concept under study than actually belongs to it (page x).

Rosenthal's concept of researcher differences is not explicit in the eight aspects listed by Hendrick but is perhaps implicit in several. I would expect, as I know a little bit about the original research team, whose work I replicated, that the expectations we each held at the start of the research may have been similar, but our philosophies of research and education were most likely different as we came from disciplines that hold different views about qualitative research. Within the realm of social science research, the continuum of values for quantitative and qualitative methods produces multiple interpretations of what replication means.

Using Hendrick's (1990) eight aspects to consider in a replication as a guide, it is a wonder that any research in the social sciences would ever be considered a true replication. This being the case, as Bornstein (1990) says, our 'exclusive reliance on imprecise replication serves to protect the professional reputation of the original investigator' (page 74).

Alternative views about replication

1. Only 'falsification' is conclusive

Popper would argue that only the falsification of a proposition, through empirical processes, is conclusive. Many recall the white/black swan analogy. Falsification requires attempted replication. 'Some propositions are more reliable than others because over time they have stood up better to disciplined attempts to refute them' (Hutcheon, 1995, page 4 of 11). This view of replication values only 'failed efforts to prove them [propositions] false' as corroborative (Hutcheon, 1995, page 4 of 11), and requires exact replication to be meaningful. As Collins (1992) stated in 1985, in pure replication 'the second experiment would amount to no more than reading the first experimental report for a second time' (page 34) and therefore would have no added value.

2. Multiple replications by multiple researchers

McKelvie (1990) believes that 'no single study can simultaneously control all extraneous variables.

Therefore repeated replication by independent investigators is the only guarantee that a phenomenon is robust' (page 83). Different studies by independent researchers, controlling different extraneous variables, might not be viewed by many as replication. At a minimum, that type of research would fall into Hendrick's (1990) category of conceptual replication, 'an attempt to convey the same crucial structure of information in the independent variables to subjects, but by a radical transformation of the procedural variables' (page 45). Hendrick believes this type of research to be high-risk. When the results are equivalent there is increased confidence in the original experiment, but if the replication fails the study is practically worthless. Collins (1992) had presented a similar view earlier saying: 'a confirmation, if it is to be worth anything in its own right, must be done in an elegant new way or in a manner that will noticeably advance the state of the art' (page 19).

Rosenbaum (2001) agrees in principal, saying, 'The epidemiologist's alternative to exact replication is the consistency of a result in a variety of repeated tests', a consistency 'not dislodged in the face of diversity' (page 82). Qualitative researchers might call this 'searching for the negative case', and it has a relationship to Popper's falsification argument. Rosenbaum believes that studies designed with applications that are varied enough to eliminate one or more of the rival explanations for the original findings are particularly valuable.

3. Achieving a degree of replication

Tsang and Kwan (1999) believe that we cannot create a scientific laboratory in the real world because the world is constantly changing and learning. In other words, exact precision is not possible in the real world. Regardless, they conclude that the fallibility of social science research makes replication even more important, even if 'the achievability of replication is a matter of degree' (page 763).

They present labels or degrees of replication (page 766) to describe different replication activities. Each cell in the chart would have a different degree of precision of replication with #1 (having same data set and same measurement and analysis) having greatest precision, and #6 generalization and extension least precision (see Table 3).

Table 3. Degrees of replication

	Same measurement and analysis	**Different measurement and analysis**
Same data set	1. Checking the analysis	2. Re-analyzing the data
Same population	3. Exact replication	4. Conceptual extension
Different population	5. Empirical generalization	6. Generalization and extension

4. 'Good enough' replication

Sing *et al* (2003) devised a reconceptualization of replication. They propose planning and conducting internal replications at the time of the original study as well as accepting the concept of 'good enough' replications. Good enough can be applied if

> it supports or verifies an earlier study with significant theoretical contributions and implications where a strict replication may not be feasible; it employs identical dependent and key independent variables (though measurements may differ) at the same level of analysis; it studies the same phenomenon with an arguably superior research design and different population. (page 539)

They also introduced the concept of 'ex-post identification'. It is possible that there are published studies, which could be viewed as 'good enough' replications of other studies in the same field, even though replication was not the original intent. Using

Table 4. Focusing replication research

	Degree of theory development	
Degree of methodological development	**Limited**	**Substantial**
Limited	e.g. knowledge management [an emerging field] Action: checking analysis and data, exact replication, empirical generalizations and conceptual extension Focus: prediction and verification for theoretical dev.	e.g. resource-based view of firm [substantial theory dev. but has not progressed adequately in methodology dev.] Action: exact replications and empirical generalizations to improve methodological aspects. Focus: verification and falsification
Substantial	e.g. strategic groups [some progress but inadequate theory advancement] Action: conceptual extension and generalization Focus: prediction, knowledge accumulation falsification for theoretical development	e.g. diversification, alliances [a field with well-developed theoretical and methodological foundations] Action: generalization and extensions Focus: broaden understanding through interdisciplinary replications, unique methods or radically different data

meta-analysis of key papers and defining research streams, supporting and non-supporting replications are identified. Not all researchers, obviously, are going to agree with 'good enough'. In an editorial commenting on meta-analysis, Steinthal (1994) stated that constructs and measures used in the studies included in a meta-analysis should be identical, although this is seldom achieved.

Sing *et al* (2003) want to encourage replications that seek to verify, especially in emerging fields, and that seek to extend or falsify basic theory because they believe that this will have the largest impact. Table 4 is their concept for focusing replication research in management theory (page 543).

4. 'Ambivalence' towards replication

Some researchers in the interpretive or constructivist paradigm of critical theory believe that their qualitative methodologies 'are now well accepted as respected ways of developing in-depth understandings about phenomenon of interest to nursing' (Berman *et al*, 1998, page 1). The validity or comprehensiveness of their findings is enhanced by combining research methods or 'methodological triangulation'. These researchers state that the principles for conducting research evolve from and operationalize the paradigm assumptions.

The following four areas would be affected by such assumptions: the relationship between the researcher and the subjects; the epistemological assumptions about the nature of knowledge and who are legitimate 'knowers'; the extent to which subjective meanings are valued and incorporated into the process of analysis and dissemination of result; and the design and conduct of the research. I don't believe that, under these research conditions, replication is an option.

Tsang and Kwan (1999) pick up on the critical theorists' 'ambivalent attitude towards replication', describing those researchers 'who contend that the principle of replicability should not be imposed on them as "overreacting" and "in danger of landing on relativism, which denies the possibility of objective truth"' (page 701).

Lately a new, yet old, methodological term has re-emerged — MOPS, Agar's (2004) acronym for multiple overdetermination patterns. Agar's definition is 'different data sources — both serendipitous and sought after — are used to build these new conceptual schemes and to test them against additional data' (page 103). I found overdetermination patterns in a neuroscience research article by Bechtel (2002). He credits his definition 'to align multi-experimental procedures to produce converging results' to Campbell and Stanley's work in the early 1960s. Originally overdetermination looked at multiple causes for the same phenomenon. These triangulation strategies can be used to support a conceptual finding, but they are not replications of any degree.

Probability

It was Raman's 1994 article that reminded me that in our behavioural research we need not have an all-or-none approach. That's why we have the concept of probability or 'the degree of belief in a theory' (page 634). He concluded that 'as we accumulate a history of successful replications, the more confidently we expect an additional successful replication but our confidence increases at a decreasing rate' as the number of replications pile up (page 637). This was expanded upon by Wells (1993) when he asked, 'How large must the number of corroborating instances be in order to conclude that a theory is adequately established?' His conclusion — there is no general answer. That didn't help much. Wells believes that we should consider the following: who did the original study? is it highly 'falsifiable'? and how much would it cost to perform another replication? when we are deciding if we have enough evidence. The quality of the researcher is a very subjective measure.

Pyett (2003), while discussing the importance of the quality of the researcher in determining the validity of a study, points out that in qualitative research the 'theoretical position, interests and political perspective … is acknowledged and even celebrated' (page 1172). In her article, she describes her personal value of the researcher's characteristics and training, her following of guidelines in her own research, and her methodological rigor; and yet she states that she still needs to 'have recourse to some test of reliability or validity' (page 1172).

Incentives for replication research

If replication is vital to the validity of research findings, why do we see so little of it in the social science research literature? Bornstein (1990) is harsh when it comes to unreplicated research.

> Science is a political as well as an intellectual enterprise. It is clear that scientists are as likely as anyone else to use illogical — even circular — reasoning in evaluating empirical data. They distort scientific information — consciously or unconsciously — to render that information consistent with their *a priori* beliefs. (page 71).

If replication is vital to the validity of research findings, why do we see so little of it in the social science research literature?

He goes on to explain that social science researchers are reluctant to conduct replication studies because of the pressure to be original. If the replication study proves the theory of the original study, publishers take a 'ho-hum, so what?' attitude. We already knew it. If the replication study does not support the original research, then the results are insignificant and meaningless based on differences in any of the aspects described previously.

Publishers and grant providers are interested in new and original research. As previously suggested, researchers have found very little academic reward in replication studies. A replication might be funded or published if the original research was controversial, either a counter-attitudinal or counter-intuitive study. Collins (1992) describes research which 'attracted enough attention to cause other scientists to criticize and later to repeat' (p.114) and concludes emphatically that 'to question the results of a passage of scientific work amounts to an accusation. There is no middle way' (p.150). Bornstein added that editors and reviewers view failures to duplicate more positively than successful replications (p.75).

In the quantitative, bench-science research literature, there is much discussion and calculation of how many times any particular study must be replicated to be viewed as valid. This also leads to discussion of the 'file-drawer syndrome', where replications that do not support a theory never see the light of day. Do we get a warped perception from the replications published? When endeavouring to practice from an evidence base, it becomes difficult to know if all of the supportive and non-supportive studies are available in the literature. There may not be a similar problem in qualitative research because it is not often that two independent researchers attempt to come to the same conclusions from the same data.

Conclusions

Have I answered the question, posed in my title? I believe so. The value of replication is very great in theory development, whether it supports the tested theory or, more importantly perhaps, not. The problem is that it is seldom attempted because it is difficult to successfully accomplish and it carries more risk than potential reward for both the replicator and the originator of the research. Interestingly, most replications attempted are not 'pure' (i.e. partial) and therefore both researchers are 'off the hook' if their findings diverge because reasons beyond their control can always be identified as confounding factors. Collins describes this phenomena in great detail, through three case studies, supporting his belief of 'science as a cultural activity rather than a locus of certain knowledge' (1992, page 1). Qualitative researchers have developed tools, replacing replication, to add rigor, legitimacy and trustworthiness to their work. Methodological triangulation, meta-analysis and MOPS are examples, but they do not appear to satisfy the reproductability need of many researchers and scholars.

In my case, the replication that I undertook was only partial but the outcomes supported the original concept of using the cognitive presence tool to measure the levels of critical thinking demonstrated in computer conferencing. The act of implementing the replication was educational and led to many questions about the use of the coding tool and definitions of terms, for which answers could not be found in the literature, or in conversation with the original researchers. As well, while the original researchers felt that their study was useful in assessing the practical value of the tool, they concluded that it was of marginal value because of the difficulties with inter-rater reliability and the inability to link the outcomes of the levels of cognitive presence to other factors, such as student outcomes in the courses. We also had some inter-rater reliability issues stemming from unclear interpretations of the conceptual definitions.

I believe our additional work on the tool added substantively to the conceptual development. The coders for my replication were graduate students in the program from which the conference data was extracted. They, and I in my teaching role, immediately saw serendipitous practical value from implementing the study. Merely identifying the levels of cognitive presentation by students in course postings gave us insight into two common issues. The interpretation of student postings in conferences by faculty who are trying to grade participation is always difficult, and faculty are looking for strategies to promote more meaningful postings by students. These inductive findings from a deductive study will lead to further exploration and to professional development of faculty and orientation for students.

I found the exercise of attempting to replicate the work of another researcher to be stimulating and valuable in many ways. My replication will be used to extend our understanding of the utility of the methodology in question and the concepts under investigation. I would recommend that researchers attempt to replicate the work of major researchers in a conceptual field before beginning an extension of that work, particularly if they are not in frequent face-to-face discussion about interpretation of meaning with the original researcher. Research built on an understanding of the original study is closer to replication than most meta-analysis.

References

M Agar (2004), 'Know when to hold 'em, know when to fold 'em: qualitative thinking outside of the university', *Qualitative Health Research*, 14(1), pages 100–112.

Y Amir (1990), 'Replication research: a 'must' for the scientific advancement of psychology', in Neuliep, *Handbook*, pages 51–70.

W Bechtel (2002), 'Aligning multiple research techniques in cognitive neuroscience: why is it important?', *Philosophy of Science,* 69, S48–S58.

H Berman, M Ford-Gilboe and J C Campbell (1998), 'Combining stories and numbers: a methodologic approach for a critical nursing sciences', *Advances in Nursing Science,* 211, pages 1–15.

R F Bornstein (1990), 'Publication, politics, experimenter bias and the replication process in social science research' in Neuliep, *Handbook*, pages 71–82.

H M Collins (1992), *Changing Order: Replication and Induction in Scientific Practice* (University of Chicago Press, Chicago).

R W Easley, C S Madden and M G Dunn (2000), 'Conducting marketing science: the role of replication in the research processes', *Journal of Business Research*, 48, pages 83–92.

R D Garrison, T Anderson and W Archer (2001), 'Critical thinking, cognitive presence and computer conferencing in distance education', *American Journal of Distance Education*, 15(1), pages 1–18.

C Hendrick (1990), 'Replications, strict replications and conceptual replications: are they important?', in Neuliep, *Handbook*, pages 45–48.

P D Hutcheon (1995), 'Popper and Kuhn on the evolution of science, *Brock Review*, 4(1/2), pages 28–37.

P A Lamal (1990), 'On the importance of replication', in Neuliep, *Handbook*, pages 31–36.

S McKelvie (1990), 'Personal comment on replications', in Neuliep, *Handbook*, pages 83 –84.

J M Morse and C Mitcham (2002), 'Exploring qualitatively derived concepts: inductive-deductive pitfalls', *International Journal of Qualitative Methods*, 1(4).

J W Neuliep (editor) (1990), *Handbook of Replication Research in the Behavioural Social Sciences* (Selected Press, Madeira, CA).

P M Pyett (2003), 'Validation of qualitative research in the 'real world'', *Qualitative Health Research*, 13(8), pages 1170–1179.

K Raman (1994), 'Inductive inference and replications: a Bayesian perspective', *Journal of Consumer Research*, 20(4), pages 633–643.

P R Rosenbaum (2001), 'Replicating effects and biases', *American Statistician*, 55(3), pages 223–227.

R Rosenthal (1990), 'Replication in behavioural research', in Neuliep, *Handbook*, pages 1–10.

J R Rossiter (2003), 'Qualifying the importance of findings', *Journal of Business Research*, 56, pages 85–88.

K Sing, S H Ang and S M Leong (2003), 'Increasing replication for knowledge accumulation in strategy research', *Journal of Management*, 29(4), pages 533–549.

B Sternthal (1994), 'Editorial', *Journal of Consumer Research*, 21.

E W K Tsang and K M Kwan (1999), 'Replication and theory development in organizational science: a critical realistic perspective', *Academy of Management Review*, 24(4), pages 759–780.

W D Wells (1993), 'Discovery-oriented consumer research', *Journal of Consumer Research*, 19, pages 489–504.