2013 Joint Conference of the 23nd International Workshop on Software Measurement (IWSM) and the Eighth International Conference on Software Process and Product Measurement (Mensura)

Worldviews, Research Methods, and their Relationship to Validity in Empirical Software Engineering Research

Kai Petersen
School of Computing
Blekinge Institute of Technology
Karlskrona, Sweden
Email: kai.petersen@bth.se

Cigdem Gencel
Facult of Computer Science
Free University of Bolzano/Bozen
Bolzano/Bozen, Italy
Email: cigdem.gencel@unibz.it

***Abstract*—Background - Validity threats should be considered and consistently reported to judge the value of an empirical software engineering research study. The relevance of specific threats for a particular research study depends on the worldview or philosophical worldview of the researchers of the study.**
Problem/Gap - In software engineering, different categorizations exist, which leads to inconsistent reporting and consideration of threats.
Contribution - In this paper, we relate different worldviews to software engineering research methods, identify generic categories for validity threats, and provide a categorization of validity threats with respect to their relevance for different world views. Thereafter, we provide a checklist aiding researchers in identifying relevant threats.
Method - Different threat categorizations and threats have been identified in literature, and are reflected on in relation to software engineering research.
Results - Software engineering is dominated by the pragmatist worldviews, and therefore use multiple methods in research. Maxwell's categorization of validity threats has been chosen as very suitable for reporting validity threats in software engineering research.
Conclusion - We recommend to follow a checklist approach, and reporting first the philosophical worldview of the researcher when doing the research, the research methods and all threats relevant, including open, reduced, and mitigated threats.

I. Introduction

Philosophical worldviews (also referred to as philosophical worldviews or paradigms) refer to "a basic set of beliefs that guide action" [1]. Four worldviews are distinguished, namely positivist (or post-positivist), constructivist (or interpretivist), advocacy/participatory, and pragmatist as below [2].

The positivists/postpositivists seek an objective reality that exists 'out there' in the world. They hold a deterministic philosophy; that is, based on careful observations and measurements, they try to make inferences to a general truth. The term positivist was replaced by post positivist after the traditional notion of the absolute truth of knowledge was challenged in the community, particularly by Popper [3] who advanced falsification and Kuhn [4] who brought a broader critique that it is not simply individual theories but whole worldviews that must occasionally shift in response to evidence. Thereafter, post positivists have taken a worldview that evidence established in research is always imperfect and fallible. Therefore, they state that they do not prove a hypothesis, but indicate a failure to reject a hypothesis.

The interpretivists seek for subjective reality, constructed by how human beings see and interpret the world in their respective context. So, truth is not absolute but relative in interpretism.

The advocacy/participatory researchers holds that research inquiry needs to be intertwined with politics and political agenda and contains an action agenda through intervention for reform that may change the lives of the participants.

The pragmatists, emphasise the research problem and using all approaches available to understand the problem, instead of focusing on the methods. The originator of this philosophy is Charles Peirce, who was influenced by Popper in many respects. The pragmatists seek a truth that is *"what is practically useful and whatever works at the time"*. Therefore, they are not committed to any philosophical view or reality, and therefore use mixed methods in their inquiries.

Which world view to adopt is also related to the 'objects of study' in a particular field; whether they are inanimate and cannot interpret themselves (as in natural sciences) or they are human beings who can interpret both themselves and their environment [5]. Therefore, for example, if a study is about people in an organization, the interpretivist world view might dominate, while if the objects of the study is a software product and some hypotheses are to be tested, a positivist world view is likely to dominate.

The worldview a researcher adopts influences which research methods to use as part of their research methodology for providing reliable evidence about the studied phenomenon of interest [6]. In empirical software engineering, we argue that the dominant worldview is pragmatism and we often use multiple methods (qualitative or quantitative) of the three world views. Therefore, it is particularly important for software engineering researchers to be explicit about their philosophical worldview and how it shaped their research to be able to argue why they chose the methods they did as well as the threats to validity of their research.

There defined a few guidelines in empirical software engineering to asses threats to validity of research results while

978-0-7695-5078-7/13 $26.00 © 2013 IEEE
DOI 10.1109/IWSM-Mensura.2013.22

using different methods. For example, Wohlin et al. [7] adopted Cook's categorisation [8] of threats to validity for experiments, while Runeson and Höst [9] defined categories and definitions for validity threats in case study research built upon [7] and [10].

However, in software engineering so far the validity threats have been defined independently of the worldviews. On the other hand, which worldview is adopted hugely affects how a researcher views the validity of a research. For example, a researcher primarily following the positivist worldview might reject conclusions of a research conducted within a research community having different worldview, as the worldview determines how we evaluate concrete studies. Hence, the relevance and value of the different worldviews has to be made explicit and validity issues of software engineering research have to be seen in the context of worldviews. Furthermore, we also identified that the established categories and definitions of validity threats is incomplete and there are some inconsistencies among them.

In this paper, we aim to clarify validity threats in empirical software engineering linking research methods to different worldviews, and consequently help in evaluating research through mirrors of the different worldviews. Furthermore, by defining a complementary classification for validity threats in software engineering considering all worldviews, and linking concrete validity threats to specific methods, we aid researchers in identifying threats relevant to their cases. For that purpose the following sub-contributions are made:

- Identify validity classifications and their meanings for the different worldviews, and compare the classifications between the views,
- Relate the validity classifications to concrete validity threats for the different worldviews, and in relation to main phases of research,
- Adopt a generic classification for validity threats considering pragmatism as the dominating philosophy in empirical software engineering research.

The paper is structured as follows: In Section II, we present different categories and definitions for validity threats in relation to different worldviews, and compare to the ones commonly used in empirical software engineering research. In Section III, we rate specific types of validity threats under each validity threat category to phases of research to guide researchers. In Section IV, we discuss the implications of this research on evaluating validity of empirical research and how researchers having different worldviews can use the results of our study in designing and communicating their research results. And finally, in Section V, we conclude our study providing future research directions.

II. Worldviews, Research Methods and Relevant Types of Validity Threats

In Table I, we provide an overview of the different worldviews with respect to their notion of truth, example research methods used, and nature of data collected about an account.

TABLE I. Research Methods and Nature of Data in Different Worldviews

World views	Truth	Research method	Data collection
Positivist/ Postpositivist	Objective (independent from participants)	Controlled experiment, case study, and survey	Quantitative (Random sampling)
Constructivist/ Interpretivist	Subjective (constructed by participants)	Case study, survey, interview	Quantitative and Qualitative (Random and purposeful)
Advocacy/ Participatory	Subjective (constructed by participants and the observer)	Action research (intervention driven)	Quantitative and Qualitative (Random and purposeful)
Pragmatist	Depends on what works at the time	Multi method research	Quantitative and Qualitative (Random and purposeful)

For example, positivist case studies test hypotheses for a concrete real world case based on measured data, see [11] for an example of an objectivist case study. On the other hand, a good example for interpretivist case study is where we investigated the perceived advantages and disadvantages of agile software development in a large company, where practitioners provide their views based on their experience and context [12]. Advocacy/participatory researchers use various methods, but action research studies are the dominating ones in the area of software engineering and can be found in the following references: [13], [14].

An example of pragmatist multi-method research is the evidence-based software engineering process [15], consisting of the steps: (1) identifying the need of information (evidence); (2) tracking down the evidence needed and critically appraise it; (3) critically reflect on the evidence provided with respect to the problem and context that the evidence should help to solve. Each step can be conducted using combinations of multiple methods. As an example, we used case study, systematic review, and value stream analysis for step 1, 2, and 3, respectively to analyze an automotive testing process [16].

Researchers working in different research environments where different worldviews dominate, define their own categorization for validity threats in relation to the research methods they use. However, in empirical software engineering in particular as, being pragmatists, we often use multi methods and this creates challenges in reporting threats to validity of our research with respect to various categories described, as well as a lot of confusion in evaluating research.

Maxwell [17] states that he does not believe that validity threats in quantitative and qualitative approaches are incompatible, and adds that rather there are important similarities between the two. He asserts that validity in a broad sense pertains a relationship between an account and something outside of this account whether this something is construed as objective reality, the constructions of participants, or a variety of other possible ways.

By following Maxwell's stance and using his categories for validity threats, below we will elaborate on the validity threats defined in empirical software engineering in relation to different worldviews.

Table II provides an overview of the validity threat categories with respect to the different world views in software engineering as well as their mapping. Threats in one row are

TABLE II. CATEGORIZATION OF THREATS MAPPING WITH RESPECT TO WORLD VIEWS

Category	Positivist (Cook [8], Wohlin et al. [7]	Interpretivist (Lincoln and Guba [18])	Advocacy/ Participatory (Greenwood and Levin [19])	Pragmatist (Runeson and Hoest [9])	Maxwell [17]
C1	Internal Validity	–	Uncontrollability	Internal Validity	Theoretical Validity
C2	External Validity	Transferability	Transcontextual	External Validity	External Generalizability
C3	Construct Validity	–	Contingency	Construct Validity	Theoretical Validity
C4	Conclusion Validity	–	–	–	Internal Generalizability
C5	–	Credibility	–	–	Descriptive Validity
C6	–	Confirmability	Subjectivity	Reliability	Interpretive Validity

similar in their meaning even though different terminologies are used in different world views.

A complete list of definitions of the different types of Validity is provided in Table III.

Category C1: The validity threats in this category deal with factors that might affect cause and effect relationships, but is unknown to the researcher. As an example, two groups of students with completely different levels of expertise use two different inspection techniques. We would like to know which inspection technique works better, but the different experience levels also largely affect the outcome, and has not been controlled for. In the case of the positivist worldview (including positivist case studies [9]) and experimental studies [7] the cause-effect refers to a statistically established relationship. However, in the cases of the interpretivist and critical theory worldview non-statistical inferences about data qualitative data are made, hence the threat still applies to them. This is also the reason for the threat being raised for all worldviews, and validity threat discussions reported in SE literature. Thus, it is not justified to exclude the threat category in research planning with the argument that no statistical cause-effect relationships are to be established.

Category C2: The validity threats in this category are concerned with the ability to generalize the results. Positivist studies aim at finding objective truths, which implies that they are valid for the respective sample (i.e. positivist studies are based on random sampling and clear definition of populations). Hence, positivists are after being able to generalize from a sample to a population. As an example, if we conduct a survey on the Swedish software industry, but have too few answers for a particular software engineering domain (e.g. embedded systems), then our sample might not be generalizable to the population at large. In interpretivist and action research studies generalizability has to be viewed differently. As Yin [10] points out, one should not talk about a sample of cases, given that one would not aim to generalize to a population. Instead, one would like to generalize to similar contexts, and find supporting cases and conflicting cases for theories, and by doing that being able to conduct cross-case comparison. From our experience of working with industry, the practitioners are often most interested in cases, in particular those that match their company context (e.g. in terms of domain, size and complexity of development, use of similar software processes, and so forth). Hence, reporting context is very important to know which cases to compare (cf. [21]). Overall, this makes clear that case studies should not be rejected due to that they represent only a single case, each case is an important contribution to learning, in particular as case studies provide a deep understanding [10], [9] of a situation in a particular context. The same applies to action research, which is also focused on being conducted in the real world, and hence produces context dependent results [20].

Category C3: The validity threats in this category are concerned with whether we measured (e.g. quantitative measurement instrument) or captured (e.g. qualitatively in an interview), what we intend to in relation to our hypothesis or theory to test. This threat is equally relevant in all world views (see Table II)

Category C4: The validity threats in this category deal with the degree to which conclusions/inferences we draw (e.g. about relationships between variables, or based on qualitative data) are reasonable. Typically this concerns if there is a statistically significant effect on the outcome for experiments. As an example, if we establish a correlation between two variables based on a very small sample, is it then reasonable to draw the conclusion that the variables are related? In a qualitative situation, we might observe in the field that developers being unmotivated while being under constant pressure through frequent releases (overload). Is it reasonable to conclude based on the number of observations that lack of unmotivation is related to constant pressure? This threat is not highlighted by all views, however, we believe that one should always consider threats in relation to whether the conclusions drawn are reasonable with respect to the collected data.

Category C5: This category refers to validity threats related to factual accuracy of the account; that is, the researchers are not making up or distorting the things they observed and it is expected to produce descriptively same accounts/data for the same event or situation. One important comment of Maxwell [17] is that this validity concerns also issues of omission (no account can include everything, and we include/exclude/omit depending on the implicit theory we have.).

Category C6: Validity threats under this category concerns with whether the inferences/conclusions follow from the account (data), not biased by the researchers during analysis. Maxwell [17] claims that this category is not so much relevant to quantitative studies, but to qualitative studies as the methods used are more apt to this types of threat when the researchers make interpretations about the data.

There are also other validity threats categories defined in the literature such as Reliability [10] (in positivist case studies) or Dependability [18] in qualitative research. These actually concern with repeatability or reproducability in research (i.e., whether we would obtain the same results if we could observe the same thing twice). For these threats, we agree with Maxwell's point of view [17] that they do not refer to an aspect of validity or separate issue of validity, but a particular type

TABLE III. Validity Threat Definitions in Different Worldviews with Respect to Categories

Category C1	**Definitions of Mapping Validity Threat Categories**
Internal validity	When the researcher is investigating whether one factor affects an investigated factor there is a risk that the investigated factor is also affected by a third factor. If the researcher is not aware of the third factor and/or does not know to what extent it affects the investigated factor, there is a threat to the internal validity. [7], [9]
Uncontrollability	The researcher does not usually have full control over that environment [20]
Theoretical validity	Refers to an accounts function as an explanation (as well as a description or interpretation, of the phenomena); that is, as a theory of some phenomenon. Any theory has two components: the concepts or categories that the theory employs, and the relationships that are thought to exist among these concepts [17]
Category C2	
External validity	This aspect of validity is concerned with to what extent it is possible to generalize the findings, and to what extent the findings are of interest to other people outside the investigated case. During analysis of external validity, the researcher tries to analyze to what extent the findings are of relevance for other cases. [7], [9]
Transferability	Transferability refers to the degree to which the results of qualitative research can be generalized or transferred to other contexts or settings. From a qualitative perspective transferability is primarily the responsibility of the one doing the generalizing. [18]
Transcontextual	Research outcome is intersection of environmental conditions, a group of people, and a variety of historical events, including the actions of participants [19]
External Generalizability	Refers to the extent to which one can extend the account of a particular situation or population to other other communities, groups, or institutions. [17]
Category C3	
Construct Validity	Refers to what extent the operational measures that are studied really represent what the researcher have in mind and what is investigated according to the research questions. [9]
Contingency	Inherent obstacles to isolation of evidence related to particular effects and constructs from the contextual "glue" in which they are naturally found, given a vast amount of shallow information. [20]
Theoretical validity	Refers to an accounts function as an explanation (as well as a description or interpretation, of the phenomena); that is, as a theory of some phenomenon. [17]
Category C4	
Conclusion validity	Focus on how sure we can be that the treatment we used in an experiment really is related to the actual outcome we observed; that is whether the results are statistically significant [7]
Internal Generalizability	Generalizing within the community, group, or institution studied to persons, events, and settings that were not directly observed or interviewed [17]
Category C5	
Credibility	The credibility criteria involves establishing that the results of qualitative research are credible or believable from the perspective of the participant in the research. Since from this perspective, the purpose of qualitative research is to describe or understand the phenomena of interest from the participant's eyes, the participants are the only ones who can legitimately judge the credibility of the results. [18]
Descriptive Validity	Factual accuracy of the account that is, researchers are not making up or distorting the things they observed (interpretivist cognition – see, hear, smell etc where positivists measure using measurement instruments). [17]
Category C6	
Confirmability	Confirmability refers to the degree to which the results could be confirmed or corroborated by others. [18]
Subjectivity	The deep involvement of researchers with client organizations in Action Research studies may hinder good research by introducing personal biases in the conclusions. [20]
Reliability	Concerns to what extent the data and the analysis are dependent on the specific researchers. Hypothetically, if another researcher later on conducted the same study, the result should be the same. [9]

of validity that arise in relation a number of validity threat categories defined above.

Based on the existing classifications of validity threats, we believe that Maxwell's classification fits best to software engineering as we being pragmatists use multi methods, and therefore require a more generic classification to cover different methods of all worldviews. Even though Maxwell provided his definitions for qualitative research, the different threats can also be applied to quantitative research, as will be illustrated in the following section by looking at individual threats and classifying them according to the categories.

Maxwell's classification has been chosen for two reasons: 1) The threat categories are defined generic enough to capture the different philosophical worldviews, i.e. they can apply to more quantitative, as well as qualitative research. We demonstrate this by categorizing threats of different worldviews according to the categories and comparing to those by Wohlin et al. [7] for experiments and Runeson and Höst [9] for case studies, and 2) The terminology of the threat categories is more intuitive, with terms such as theoretical validity, interpretative validity, and descriptive validity, and generalisability.

Figure 1 provides the classification of threats according to the definition of Maxwell. As can be seen from Table II all four validity threats raised by Wohlin et al. [7] for experiments and by Greenwood and Levn [19] for action research are covered by Maxwell [17]. As for the categories by [9], only reliability is not included due to the reasons we just stated. Similarly, dependability [18] is not included. In addition, Maxwell added two more categories: descriptive validity and interpretive validity in his categories. We argue, as indicated by the figure, that repeatability (or dependability or reproducibility) follows from addressing the threats defined by Maxwell. For example:

- If we do not have a means to draw conclusions from the data (interpretative validity), we will very likely draw different conclusions assuming we could repeat the research.
- If we are not aware of generalizability, we can not repeat the study in different contexts and compare (e.g. due to not knowing about the context)
- If we do not have means to collect correct data, we are likely to get different results when measuring the same attribute.

After having identified the categories for validity threats, we map a selection of existing and well known threats in relation to world views and a generic research process for software engineering, consisting of (1) design of study and data collection, and (2) analyze data.

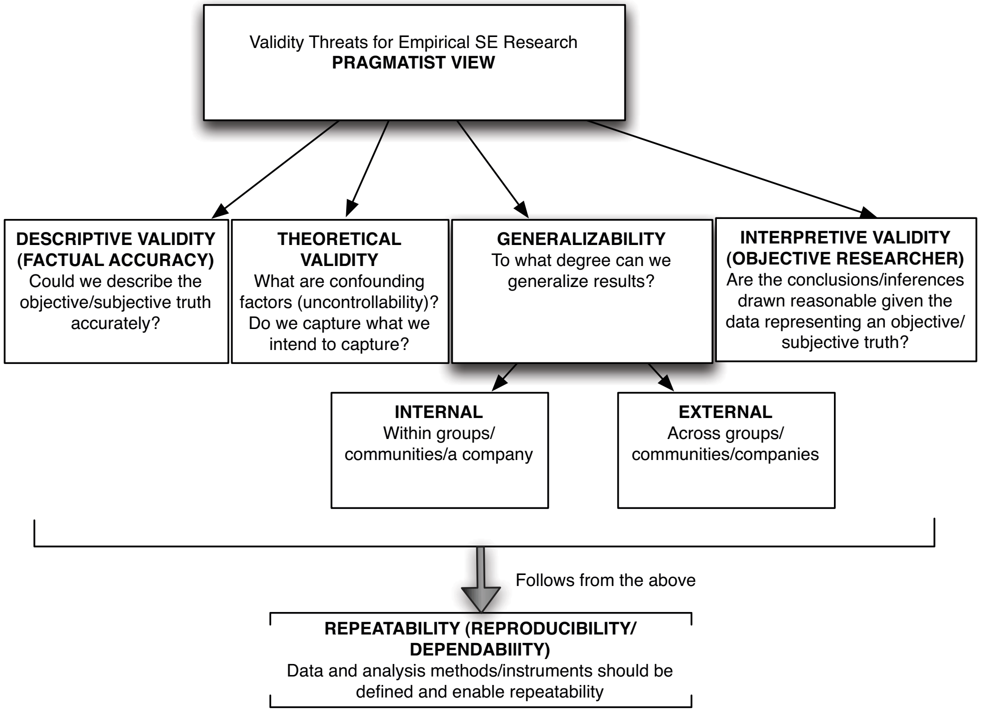
Validity Threats for Empirical SE Research
PRAGMATIST VIEW
DESCRIPTIVE VALIDITY (FACTUAL ACCURACY)
Could we describe the objective/subjective truth accurately?
THEORETICAL VALIDITY
What are confounding factors (uncontrollability)? Do we capture what we intend to capture?
GENERALIZABILITY
To what degree can we generalize results?
INTERPRETIVE VALIDITY (OBJECTIVE RESEARCHER)
Are the conclusions/inferences drawn reasonable given the data representing an objective/subjective truth?
INTERNAL
Within groups/communities/a company
EXTERNAL
Across groups/communities/companies
Follows from the above
REPEATABILITY (REPRODUCIBILITY/ DEPENDABIIITY)
Data and analysis methods/instruments should be defined and enable repeatability

Fig. 1. Categorization of Validity Threats (Pragmatic Software Engineering View)

III. Validity Threats in Relation to Software Engineering Research Phases

We distinguish two main phases in research to identify relevant validity threats (1) data collection, and (2) data analysis. Design of data collection and analysis procedures are considered under each phase to avoid confusion. That is, generalizability, theoretical validity, and descriptive validity are relevant and have to be considered both in the design of the data collection procedures as well as during the collection of the data. Then, when the data is collected, interpretative validity and generalisability comes into play in the analysis of the results.

Figure 2 provides an overview of the research process. It shows the different philosophical worldviews, and concrete related validity threats. The threats are classified according to the threat categories provided in Figure 1. The actual threats are identified based on literature of research methods representing the different worldviews (see e.g. [7], [19], [20]). It should be highlighted that we do not aim at providing a complete list of threats, but rather document the threats to make the difference between the worldviews explicit based on the individual threats, even though the main validity threat categories can be shared between them.

Threats in the positivist worldview: We first compare the positivist view with the others to highlight what is unique about it. The threats of the positivist view are all related to quantitative measures used in hypotheses testing. Therefore, reliability of measures and reliability of treatment implementation are of particular relevance. The validity of results while measuring and comparing can, however, be influenced by various factors (e.g. selection of subjects, history, maturation, or researcher expectations). The main difference to the other views is how generalizability and interpretation are viewed. Generalizability here is focused on generalizing the sample to the defined population, while in the other world views sampling has to be purposeful, for example, we should choose a study setting that provides interesting results, not just repeating existing ones (saturation), e.g. by studying a different domain for the same object of study (let us say agile software development). Furthermore, interpretative validity is different due to that the conclusions follow from the results of the statistical analysis, hence statistical power and assumptions of statistical tests are emphasized over researcher bias, or the researchers' contextual blindness.

Threats in the interpretivist worldview: Some threats in the interpretivist view are shared with the objectivist view. For example, constructs also have to be well defined to be able to compare different cases, or ask the right interview questions. On the other hand, in the objectivist view construct definition helps defining the right measures. Furthermore, evaluation comprehension is important in both cases. A subject in a positivist experiment might not like to be evaluated on how well he or she can inspect a document, while in the interpretivist case a subject might not answer honestly based on questions regarding his or her skills in an interview. In relation to the positivist view, descriptive validity has to be highlighted. Given the vast amount of qualitative information gathered, the threat is more significant, and it is more challenging to capture all relevant information (e.g. in observations or interviews

World Views

	Positivist	Interpretivist	Participatory (Critical Theory)	Pragmatist
Data Collection (record, store, review, and revise)	**Theoretical Validity** - Reliability of measures - Reliability of treatment implementation - Heterogeneity of subjects - Constructs (e.g. theory) not well defined, unclear - Mono-method bias (use single measure) - Evaluation apprehension - Selection of subjects - Ambiguity about direction of causal influence - History affects results - Maturation (behavior change over time) - Poor instrumentation for data collection - Compensatory rivalry between subjects with different treatments - Researcher expectations **Generalizability** - Interaction of selection and treatment - Interaction of setting and treatment - Interaction of history and treatment	**Theoretical Validity** - Constructs (e.g. theory, object of study) not well defined, unclear - Mono-method bias - Evaluation apprehension - Selection of subjects - Ambiguity about direction of inferences derived from data - History affects results - Maturation (behavior change over time) - Poor instrumentation for data collection **Generalizability** - Saturation, i.e. non-purposeful selection study context/questions **Descriptive Validity** - poor recording of data - too many steps of interpretation	**Theoretical Validity** - Constructs (e.g. theory, object of study) not well defined, unclear - Mono-method bias - Descriptive accuracy - Selection of subjects - Ambiguity about direction of inferences derived from data - History affects results - Maturation (behavior change over time) - Learning effects - Poor instrumentation for data collection - political and social threats - Resistance of change - few iterations **Generalizability** - Saturation, i.e. non-purposeful selection study context/questions **Descriptive Validity** - poor recording of data - too many steps of interpretation	Combinations of the views, depending on how research stances are combined (multi-method)
Data Analysis (interpret, report, review, and revise)	**Interpretative Validity** - Low stat. power - Invalid assumptions of stat. tests	**Interpretative Validity** - Researcher bias in drawing conclusions **Generalizability** - Lack of context definition and awareness in interpretation	**Interpretative Validity** - Researcher bias in drawing conclusions - Researcher suffers from organizational/contextual blindness **Generalizability** - Lack of context definition and awareness in interpretaiton	Combinations of the views, depending on how research stances are combined (multi-method)

Fig. 2. World Views, Research Process, and Validity Threats

without recording). Also, the interpretation of field data in interpretivist studies might be influenced by researcher bias. Even though the interpretivist worldview highlights subjective truth, this only refers to the truth of the subjects being studied. The researcher should provide an objective interpretation of that subjective truth.

Threats in the critical theory worldview: Critical theory is in some sense similar to the interpretivist view considering the threats. Both are, e.g., similar in relation to how they ought to treat external validity, given that both types of studies are embedded in a specific research context, and both rely on gathering qualitative as well as quantitative information. The main difference is that the critical theory worldview aims freeing people from their restrictive thoughts, by introducing interventions. Action research is an intervention driven methodology, that also relies on the researcher being embedded and engaged with the research setting over a long period of time, being "part of the team". Furthermore, the researcher introduces interventions that are to be evaluated in multiple cycles. This raises some unique threats, such as political and social threats due to the very close engagement of the researcher in the research setting [19], and the number of iterations (also referred to as action research cycles) being of importance. In interpretative validity, we add that the researcher suffers from organizational and contextual blindness due to the long-term engagement as well. The validity threats are highlighted by the following definition of what makes a valid action research, according to "research outcomes are well-grounded if the focus of the inquiry, both in its parts and as a whole, is taken through as many cycles as possible, by as many group members as possible, with as many individual diversity as possible, and collective unity of approach as possible" [22].

Threats in the pragmatism worldview: As we elaborated before, this worldview considers truth what works at the time. This sentence could be interpreted as an alibi for conducting poor and invalid research, which of course must not be the case. In particular, what works well at the time, and in which context, has to be based on high quality research. Given that, as we argued before, software engineering research is following the pragmatism worldview, all threats are potentially relevant in our research field.

IV. Discussion

A. Consequences for Research Evaluation

The presence of validity threats not addressed in research designs are often reasons to provide low quality scores to studies when aggregating evidence (e.g. in systematic reviews [23]), or to reject them in the peer review process.

Given that the relevance of specific threats depends on the philosophical worldview taken in the research, it is important to take the worldview into consideration when evaluating a study. That is, a study should not be scored low (systematic review) or be rejected (peer review) for threats that are not of relevance for the worldview. A good example is the case study research method, which in some cases is rejected due to that only a single case is investigated, given that the objectivist view of sampling is applied. Though, as discussed earlier, this should be avoided given that case studies should be purposeful, rather than representative of a large population. A reason to reject a case might, for instance, be saturation (i.e. no new insights are gained). A pre-requisite for fair evaluations is hence an awareness of the different worldviews with respect to validity threats. We provided this through definitions of the worldviews, examples of software measurement research studies relating to them, and the identification of definitions of threat categories and concrete threats in relation to world views.

B. How to Use the Results of this Paper

We saw a need of clarification of the categorisation of validity threats in software engineering as the reasons for the differences between different categories were not well-described. A consequence is inconsistent reporting of the validity threats in studies, and not relating the validity threats to the world views being taken in the research. Our aim with this paper was to contribute to the clarification in three ways, namely by: (1) defining the world views, and providing examples from the software engineering measurement research area; (2) identifying an easy to understand and generic categorization of validity threats; (3) identifying concrete validity threats for each world view and making the differences and similarities explicit.

The concrete validity threats might guide researchers as a checklist depending on the worldview taken in their research. The checklist contains four questions to be answered:

1. Which of the world views you had when doing this research?

a) Positivist
b) Interpretivist
c) Advocacy/Partcipatory
d) Pragmatist

2. What type of research you are conducting (based on research questions) (see Table IV)? More detailed definitions of the types of research, as well as related research question types are defined in [24].

TABLE IV. Research Type

Method	Descriptive	Exploratory	Explanatory	Improving
Case Study	√	√	√	
Survey	√	√	√	
Interview	√	√	√	
Experiment			√	
Action research				√

3. Choose which research method(s) you used in your study (see Table V)? Note that the answer to this question should be consistent with choices on question 1. and 2.

4. Check which of the validity threat categories apply to your study for each phase. Figure 2 aids in identifying relevant threats with regard to world views. In future work, the list can be completed by reviewing literature on specific research methodologies to identify threats unique to them.

We also aim at improving the reporting of validity threats. We propose to report validity threats as follows: Validity threats for design and conduct of data collection and analysis

TABLE V. CHOICE OF RESEARCH METHOD

Method	Positivist	Interpretivist	Critical Theory	Pragmatist
Case Study	√	√		√
Survey	√			√
Interview		√		√
Experiment	√			√
Action research			√	√

should be reported separately, so that the countermeasures could be identified accordingly where they are needed in the research process. Furthermore, all threats potentially relevant should be documented. If a threat is not reported, it can mean two things: Either the threat was addressed and hence not reported, or it was overlooked. Consequently, it would be useful to report:

- Open threats that were not reduced or mitigated in the research design.
- Reduced threats that are still of relevance, but countermeasures have been taken to address them.
- Mitigated threats not of any relevance for the study, as they are completely mitigated by the design of the study, or the worldview taken.

Such reporting is not only useful for researchers, but also for practitioners who would be aided in the decision of whether to adopt a solution proposed in the studies.

C. Future Directions for Research

The work presented in this paper is a starting point to identify relevant validity threats, and to make reporting of validity threats more consistent. Further work is needed to achieve this goal. In particular, the list of threats (as provided in Figure 2) has to be further extended through the review of literature on research methods, as well as review on reported threats in empirical software engineering research studies.

It would be of interest to also assess the quality of current discussions of validity, and whether the relevant threats to validity have been discussed for the reported studies. Hence, we propose to conduct systematic literature studies on validity threat reporting in different research areas, and for different study types.

Besides identifying and improving the reporting, we also should identify countermeasures in future work. That is, when a relevant threat has been identified, research design decisions should be related to different threats, so that software engineering researchers have support in increasing the validity of their research.

V. CONCLUSION

We identified the need for consistent and complete reporting of validity threats in software engineering measurement research. In software engineering different categorizations exist in relation to research methods, but not considering the worldviews. This is particularly important for software engineering research as being pragmatists, we use multiple methods in research. A consequence is inconsistent and incomplete reporting of validity threats.

For that purpose we characterized worldviews based on literature, and provided examples of existing studies in relation to the worldviews. We identified different classifications of validity threats, and found Maxwell's classification most suitable for software engineering, given that it represents a pragmatist world view. Four categories are defined, namely interpretative validity, generalizability (internal and external), theoretical validity, and descriptive validity. The overall categories are defined abstract enough that validity of all worldviews can be related to them. This further indicates that they support the pragmatist view, which encompasses the different world views depending on what is needed.

Specific threats of validity have been identified and classified according to the different world views. The list is not complete, but rather is used to clarify the difference between threats in relation to the world views.

This paper provides a starting point to choose relevant threats and report them consistently by following the steps of: (1) choosing a world view for the research, (2) choosing the type of research, and (3) choosing the method. These choices then guide which threats to consider in the research in step (4).

In future work, the list of threats in relation to world views has to be further extended. Furthermore, existing literature has to be critically assessed with respect to validity threats, and countermeasures implemented to reduce or mitigate the threats. Such an assessment provides important learnings to further improve software engineering research.

It would also be of interest to review literature investigating which worldview is dominant in different sub-disciplines of software engineering. Systematically investigating the worldviews represented by studies would also empirically substantiate our argument that software engineering represents the pragmatist worldview.

ACKNOWLEDGMENT

This work has been supported by ELLIIT, the Strategic Area for ICT research, funded by the Swedish Government.

REFERENCES

[1] E. G. Guba, *The paradigm dialog*. SAGE Publications, Incorporated, 1990.

[2] J. W. Creswell, *Research design: qualitative, quantitative, and mixed methods approaches*, 3rd ed. Thousand Oaks, Calif.: Sage, 2009.

[3] K. R. Popper, *The logic of scientific discovery*. Psychology Press, 2002, no. 117.

[4] T. S. Kuhn, *The structure of scientific revolutions*. University of Chicago press, 2012.

[5] R. Valerdi and H. L. Davidz, "Empirical research in systems engineering: challenges and opportunities of a new frontier," *Systems Engineering*, vol. 12, no. 2, pp. 169–181, 2009.

[6] S. F. Brown, "Naivety in systems engineering research: are we putting the methodological cart before the philosophical horse," in *7th Annual Conference on Systems Engineering Research (CSER 2009)*, 2009.

[7] C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, and A. Wesslén, *Experimentation in Software Engineering*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012.

[8] T. D. Cook and D. T. Campbell, *Quasi-experimentation: design & analysis issues for field settings*. Boston: Houghton Mifflin, 1979.

[9] P. Runeson and M. Höst, "Guidelines for conducting and reporting case study research in software engineering," *Empirical Software Engineering*, vol. 14, no. 2, pp. 131–164, 2009.

[10] R. K. Yin, *Case study research: design and methods*, 4th ed. London: SAGE, 2009.

[11] P. Poba-Nzaou, L. Raymond, and B. Fabi, "Adoption and risk of erp systems in manufacturing smes: a positivist case study," *Business Process Management Journal*, vol. 14, no. 4, pp. 530–550, 2008.

[12] K. Petersen and C. Wohlin, "The effect of moving from a plan-driven to an incremental software development approach with agile practices," *Empirical Software Engineering*, vol. 15, no. 6, pp. 654–693, 2010.

[13] D. Baca and K. Petersen, "Countermeasure graphs for software security risk assessment: An action research," *Journal of Systems and Software*, 2013.

[14] J. H. Iversen, L. Mathiassen, and P. A. Nielsen, "Managing risk in software process improvement: an action research approach," *Mis Quarterly*, vol. 28, no. 3, pp. 395–433, 2004.

[15] B. A. Kitchenham, T. Dyba, and M. Jorgensen, "Evidence-based software engineering," in *Software Engineering, 2004. ICSE 2004. Proceedings. 26th International Conference on*. IEEE, 2004, pp. 273–281.

[16] A. Kasoju, K. Petersen, and M. V. Mäntylä, "Analyzing an automotive testing process with evidence-based software engineering," *Information and Software Technology, in print*, 2013.

[17] J. A. Maxwell, "Understanding and validity in qualitative research," *Harvard educational review*, vol. 62, no. 3, pp. 279–301, 1992.

[18] Y. S. Lincoln and E. G. Guba, *Naturalistic inquiry*. Beverly Hills, Calif.: Sage, 1985.

[19] D. J. Greenwood and M. Levin, *Introduction to action research: social research for social change*, 2nd ed. Thousand Oaks, Calif.: SAGE, 2007.

[20] N. Kock, "The three threats of action research: a discussion of methodological antidotes in the context of an information systems study," *Decision support systems*, vol. 37, no. 2, pp. 265–286, 2004.

[21] K. Petersen and C. Wohlin, "Context in industrial software engineering research," in *Proceedings of the 2009 3rd International Symposium on Empirical Software Engineering and Measurement*. IEEE Computer Society, 2009, pp. 401–404.

[22] K. Herr and G. L. Anderson, *The action research dissertation: a guide for students and faculty*. Thousand Oaks, Calif.: Sage, 2005.

[23] B. Kitchenham and S. Charters, "Guidelines for performing systematic literature reviews in software engineering," *Engineering*, vol. 2, no. EBSE 2007-001, 2007.

[24] S. Easterbrook, J. Singer, M.-A. Storey, and D. Damian, "Selecting empirical methods for software engineering research," in *Guide to advanced empirical software engineering*. Springer, 2008, pp. 285–311.