The Qualitative Report

Volume 10 | Number 3 Article 3

9-1-2005

The Application of Interrater Reliability as a Solidification Instrument in a Phenomenological Study

Joan F. Marques
Woodbury University, jmarques01@earthlink.net

Chester McCall
Pepperdine University, cmccall@pepperdine.edu

Recommended APA Citation

Marques, J. F., & McCall, C. (2005). The Application of Interrater Reliability as a Solidification Instrument in a Phenomenological Study. *The Qualitative Report*, *10*(3), 439-462. https://doi.org/10.46743/2160-3715/2005.1837

The Application of Interrater Reliability as a Solidification Instrument in a Phenomenological Study

Abstract

Interrater reliability has thus far not been a common application in phenomenological studies. However, once the suggestion w a s brought up by a team of supervising professors during the preliminary orals of a phenomenological study, the utilization of this verification tool turned out to be vital to the credibility level of this type of inquiry, where the researcher is perceived as the main instrument and where bias may, hence, be difficult to eliminate. With creativeness and the appropriate calculation approach the researcher of the here reviewed qualitative study managed to apply this verification tool and found that the establishment of interrater reliability served as a great solidification to the research findings.

Keywords

Phenomenology, Interrater Reliability, Applicability, Bias Reduction, Qualitative Stud y, Research Findings, and Study Solidification

The Qualitative Report Volume 10 Number 3 September 2005 439-462
http://www.nova.edu/ssss/QR/QR10-3/marques.pdf

The Application of Interrater Reliability as a Solidification Instrument in a Phenomenological Study

Joan F. Marques
Woodbury University, Burbank, California

Chester McCall
Pepperdine University, Malibu, California

Interrater reliability has thus far not been a common application in phenomenological studies. However, once the suggestion was brought up by a team of supervising professors during the preliminary orals of a phenomenological study, the utilization of this verification tool turned out to be vital to the credibility level of this type of inquiry, where the researcher is perceived as the main instrument and where bias may, hence, be difficult to eliminate. With creativeness and the appropriate calculation approach the researcher of the here reviewed qualitative study managed to apply this verification tool and found that the establishment of interrater reliability served as a great solidification to the research findings. Key Words: Phenomenology, Interrater Reliability, Applicability, Bias Reduction, Qualitative Study, Research Findings, and Study Solidification

Introduction

This paper intends to serve as support for the assertion that interrater reliability should not merely be limited to being a verification tool for quantitative research, but that it should be applied as a solidification strategy in qualitative analysis as well. This should be applied particularly in a phenomenological study, where the researcher is considered the main instrument and where, for that reason, the elimination of bias may be more difficult than in other study types.

A "verification tool," as interrater reliability is often referred to in quantitative studies, is generally perceived as a means of verifying coherence in the understanding of a certain topic, while the term "solidification strategy," as referred to in this case of a qualitative study, reaches even further: Not just as a means of verifying coherence in understanding, but at the same time a method of strengthening the findings of the entire qualitative study. The following provides clarification of the distinction in using interrater reliability as a verification tool in quantitative studies versus using this test as a solidification tool in qualitative studies. Quantitative studies, which are traditionally regarded as more scientifically based than qualitative studies, mainly apply interrater reliability as a percentage-based agreement in findings that are usually fairly straightforward in their interpretability. The interraters in a quantitative study are not necessarily required to engage deeply into the material in order to obtain an

understanding of the study's findings for rating purposes. The findings are usually obvious and require a brief review from the interraters in order to state their interpretations. The entire process can be a very concise and insignificant one, easily understandable among the interraters, due to the predominantly numerical-based nature of the quantitative findings.

However, in a qualitative study the findings are usually not represented in plain numbers. This type of study is regarded as less scientific and its findings are perceived in a more imponderable light. Applying interrater reliability in such a study requires the interraters to engage in attentive reading of the material, which then needs to be interpreted, while at the same time the interraters are expected to display a similar or basic understanding of the topic. The use of interrater reliability in these studies as more than just a verification tool because qualitative studies are thus far not unanimously considered scientifically sophisticated. It is seen more as a *solidification* tool —that can contribute to the quality of these types of studies and the level of seriousness with which they will be considered in the future. As explained earlier, the researcher is usually considered the instrument in a qualitative study. By using interrater reliability as a solidification tool, the interraters could become true validators of the findings of the qualitative study, thereby elevating the level of believability and generalizability of the outcomes of this type of study. As a clarification to the above, as the "instrument" in the study the researcher can easily fall into the trap of having his or her bias influence the study's findings. This may happen even though the study guidelines assume that he or she will dispose of all preconceived opinions before immersing himself or herself into the research. Hence, the act of involving independent interraters, who have no prior connection with the study, in the analysis of the obtained data will provide substantiation of the "instrument" and significantly reduce the chance of bias influencing the outcome. Regarding the "generalizability" enhancement Myers (2000) asserts

> Despite the many positive aspects of qualitative research, [these] studies continue to be criticized for their lack of objectivity and generalizability. The word 'generalizability' is defined as the degree to which the findings can be generalized from the study sample to the entire population. (¶ 9)

Myers continues that

> The goal of a study may be to focus on a selected contemporary phenomenon […] where in-depth descriptions would be an essential component of the process. (¶ 9)

This author subsequently suggests that, "in such situations, small qualitative studies can gain a more personal understanding of the phenomenon and the results can potentially contribute valuable knowledge to the community" (¶ 9).

It is exactly for this purpose, the potential contribution of valuable knowledge to the community, that the researcher mentioned the elevation of generalizability in qualitative studies, through the application of interrater reliability as a solidification and thus bias-reducing tool.

Before immersing into specifics it might be appropriate to explain that there are two main prerequisites considered when applying interrater reliability to qualitative research: (1) The data to be reviewed by the interraters should only be a segment of the total amount, since data in qualitative studies are usually rather substantial and interraters usually only have limited time and (2) It needs to be understood that there may be different configurations in the packaging of the themes, as listed by the various interraters, so that the researcher will need to review the context in which these themes are listed in order to determine their correspondence (Armstrong, Gosling, Weinman, & Marteau, 1997). It may also be important to emphasize here that most definitions and explanations about the use of interrater reliability to date are mainly applicable to the quantitative field, which suggests that the application of this solidification strategy in the qualitative area still needs significant review and subsequent formulation regarding its possible applicability.

This paper will first explain the two main terms to be used, namely "interrater reliability" and "phenomenology," after which the application of interrater reliability in a phenomenological study will be discussed. The phenomenological study that will be used for analysis in this paper is one that was conducted to establish a broadly acceptable definition of spirituality in the workplace. In this study the researcher interviewed six selected participants in order to obtain a listing of the vital themes of spirituality in the workplace. This process was executed as follows: First, the researcher formulated the criteria, which each participant should meet. Subsequently, she identified the participants. The six participants were selected through a snowball sampling process: Two participants referred two other participants who each referred to yet another eligible person. The researcher interviewed each participant in a similar way, using an interview protocol that was validated on its content by two recognized authors on the research topic, Drs. Ian Mitroff and Judi Neal.

- Ian Mitroff is "distinguished professor of business policy and founder of the USC Center for Crisis Management at the Marshall School of Business, University of Southern California, Los Angeles. (Ian I. Mitroff, 2005, ¶ 1). He has published "over two hundred and fifty articles and twenty-one books of which his most recent are *Smart Thinking for Difficult Times: The Art of Making Wise Decisions*, *A Spiritual Audit of Corporate America*, and *Managing Crises Before They Happen* (Ian I. Mitroff, ¶ 4).

- Judi Neal is the founder of the Association for Spirit at Work and the author of several books and "numerous academic journal articles on spirituality in the workplace" (Association for Spirit at Work, 2005, ¶ 10-11). She has also established her authority in the field of spirituality in the workplace in her position of "executive director of The Center for Spirit at Work at the University of New Haven, […] a membership organization and clearinghouse that supports personal and organizational transformation through coaching, education, research, speaking, and publications" (School of Business at the University of New Haven, 2005, ¶ 2).

After transcribing the six interviews the researcher developed a horizonalization table; all six answers to each question were listed horizontally. She subsequently eliminated redundancies in the answers and clustered the themes that emerged from this

process, which in phenomenological terms is referred to as "phenomenological reduction." This process was fairly easy, as the majority of questions in the interview protocol were worded in such a way that they solicited enumerations of topical phenomena from the participants. To clarify this with an example one of the questions was "What are some words that you consider to be crucial to a spiritual workplace?" This question solicited a listing of words that the participants considered identifiable with a spiritual workplace. From six listings of words, received from six participants, it was relatively uncomplicated to distinguish overlapping words and eliminate them. Hence, phenomenological reduction is much easier to execute these types of answers when compared to answers provided in essay-form. This, then, is how the "themes" emerged. To provide the reader with even more clarification regarding the question formulations, the interview protocol that was used in this study is included as an appendix (see Appendix A).

Interrater Reliability

Interrater reliability is the extent to which two or more individuals (coders or raters) agree. Although widely used in quantitative analyses, this verification strategy has been practically barred from qualitative studies since the 1980's because "a number of leading qualitative researchers argued that reliability and validity were terms pertaining to the quantitative paradigm and were not pertinent to qualitative inquiry" (Morse, Barrett, Mayan, Olson, & Spiers, 2002, p. 1). "Interrater reliability addresses the consistency of the implementation of a rating system" (Colorado State University, 1997, ¶ 1). The CSU on-line site further clarifies interrater reliability as follows:

> A test of interrater reliability would be the following scenario: Two or more researchers are observing a high school classroom. The class is discussing a movie that they have just viewed as a group. The researchers have a sliding rating scale (1 being most positive, 5 being most negative) with which they are rating the student's oral responses. Interrater reliability assesses the consistency of how the rating system is implemented. For example, if one researcher gives a "1" to a student response, while another researcher gives a "5," obviously the interrater reliability would be inconsistent. Interrater reliability is dependent upon the ability of two or more individuals to be consistent. Training, education and monitoring skills can enhance interrater reliability. (¶ 2)

Tashakkori and Teddlie (1998) refer to this type of reliability as "interjudge" or "interobserver," describing it as the degree to which ratings of two or more raters or observations of two or more observers are consistent with each other. According to these authors, interrater reliability can be determined by calculating the correlation between a set of ratings done by two raters ranking an attribute in a group of individuals. Tashakkori and Teddlie continue "for qualitative observations, interrater reliability is determined by evaluating the degree of agreement of two observers observing the same phenomena in the same setting" (p. 85).

In the past several years interrater reliability has rarely been used as a verification tool in qualitative studies. A variety of new criteria were introduced for the assurance of credibility in these research types instead. According to Morse et al. (2002), this was particularly the case in the United States. The main argument against using verification tools with the stringency of interrater reliability in qualitative research has, so far, been that "expecting another researcher to have the same insights from a limited data base is unrealistic" (Armstrong et al., 1997, p. 598). Many of the researchers that oppose the use of interrater reliability in qualitative analysis argue that it is practically impossible to obtain consistency in qualitative data analysis because "a qualitative account cannot be held to represent the social world, rather it 'evokes' it, which means, presumably, that different researchers would offer different evocations" (Armstrong et al., p. 598).

On the other hand, there are qualitative researchers who maintain that responsibility for reliability and validity should be reclaimed in qualitative studies, through the implementation of verification strategies that are integral and self-correcting during the conduct of inquiry itself (Morse et al., 2002). These researchers claim that the currently used verification tools for qualitative research are more of an evaluative (post hoc) than of a constructive (during the process) nature (Morse et al.), which leaves room for assumptions "that qualitative research must therefore be unreliable and invalid, lacking in rigor, and unscientific" (Morse et al., p. 4). These investigators further explain that post-hoc evaluation does "little to identify the quality of [research] decisions, the rationale behind those decisions, or the responsiveness and sensitivity of the investigator to data" (Morse et al., p. 7) and can therefore not be considered a verification strategy. The above-mentioned researchers emphasize that the currently used post-hoc procedures may very well evaluate rigor but do not ensure it (Morse et al.).

The concerns addressed by Morse et al. (2002) above about verification tools in qualitative research being more of an evaluative nature (post hoc) than of a constructive (during the process) nature can be omitted by utilizing interrater reliability as it was applied to this study, which is, right *after* the initial attainment of themes by the researcher yet *before* formulating conclusions based on the themes registered. This method of verifying the study's findings represents a constructive way (during the process) of measuring the consistency in the interpretation of the findings rather than an evaluative (post hoc) way. It therefore avoids the problem of concluding insufficient consistency in the interpretations after the study has been completed and it leaves room for the researcher to further substantiate the study before it is too late. The substantiation can happen in various ways. For instance, this might be done by seeking additional study participants, adding their answers to the material to be reviewed, performing a new cycle of phenomenological reduction, or resubmitting the package of text to the interraters for another round of theme listing.

As suggested on the Colorado State University (CSU) website (1997) interrater reliability should preferably be established outside of the context of the measurement in your study. This source claims that interrater reliability should preferably be executed as a side study or pilot study. The suggestion of executing interrater reliability as a side study corresponds with the above-presented perspective from Morse et al. (2002) that verification tools should not be executed post-hoc, but constructively during the execution of the study. As stated before, the results from establishing interrater reliability as a "side study" at a critical point during the execution of the main study (see

explanation above) will enable the researcher, in case of insufficient consistency between the interraters, to perform some additional research in order to obtain greater consensus. In the opinion of the researcher of this study, the second option suggested by CSU, using interrater reliability as a "pilot study", would mainly establish consistency in the understandability of the instrument. In this case such would be the interview protocol to be used in the research, since there would not be any findings to be evaluated at that time. However, the researcher perceives no difference between this interpretation of interrater reliability and the content validation here applied to the interview protocol by Mitroff and Neal. The researcher further questions the value of such a measurement without the additional review of study findings, or a part thereof. For this reason, the researcher decided that interrater reliability in this qualitative study would deliver optimal value if performed on critical parts of the study findings. This, then, is what was implemented in the here reviewed case.

Phenomenology

A phenomenological study entails the research of a phenomenon by obtaining authorities' verbal descriptions based on their perceptions of this phenomenon: aiming to find common themes or elements that comprise the phenomenon. The study is intended to discover and describe the elements (texture) and the underlying factors (structure) that comprise the experience of the researched phenomenon.

Phenomenology is regarded as one of the frequently used traditions in qualitative studies. According to Creswell (1998) a phenomenological study describes the meaning of the lived experiences for several individuals about a concept or the phenomenon. Blodgett-McDeavitt (1997) presents the following definition,

> Phenomenology is a research design used to study deep human experience. Not used to create new judgments or find new theories, phenomenology reduces rich descriptions of human experience to underlying, common themes, resulting in a short description in which every word accurately depicts the phenomenon as experienced by co-researchers. (¶ 10)

Creswell suggests for a phenomenological study the process of collecting information should involve primarily in-depth interviews with as many as 10 individuals. According to Creswell, "Dukes recommends studying 3 to 10 subjects, and the Riemen study included 10. The important point is to describe the meaning of a small number of individuals who have experienced the phenomenon" (p. 122).

Given these recommendations, the researcher of the phenomenological study described here chose to interview a number of participants between 3 and 10 and ended up with the voluntary choice of 6.

Creswell (1998) describes the procedure that is followed in a phenomenological approach to be undertaken:

> In a natural setting where the researcher is an instrument of data collection who gathers words or pictures, analyzes them inductively, focuses on the

> meaning of participants, and describes a process that is expressive and persuasive in language. (p. 14)

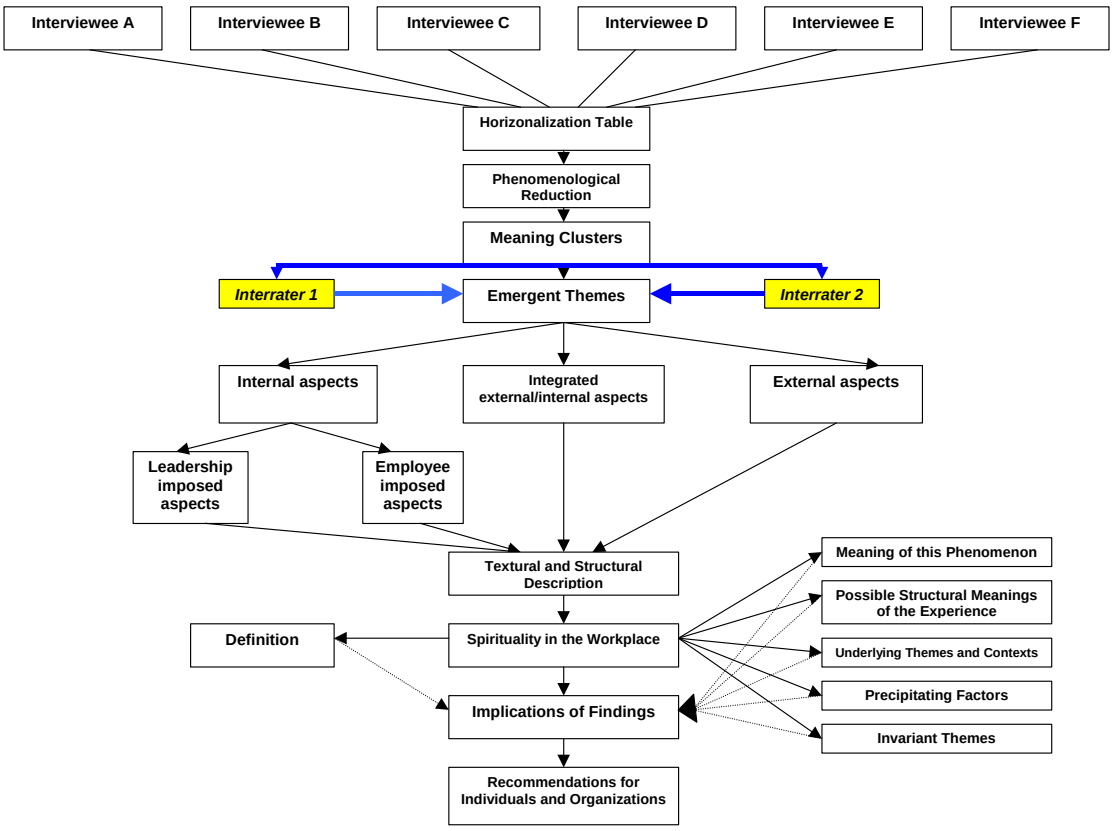
Like all qualitative studies, the researcher who engages in the phenomenological approach should realize that "phenomenology is an influential and complex philosophic tradition" (Van Manen, 2002a, ¶1) as well as "a human science method" (Van Manen, 2002a, ¶2), which "draws on many types and sources of meaning" (Van Manen, 2002b, ¶1).

Creswell (1998) presents the procedure in a phenomenological study as follows:

1. The researcher begins [the study] with a full description of his or her own experience of the phenomenon (p. 147).
2. The researcher then finds statements (in the interviews) about how individuals are experiencing the topic, lists out these significant statements (horizonalization of the data) and treats each statement as having equal worth, and works to develop a list of nonrepetitive, nonoverlapping statements (p. 147).
3. These statements are then grouped into "meaning units": the researcher lists these units, and he or she writes a description of the "textures" (textural description) of the experience - what happened - including verbatim examples (p. 150).
4. The researcher next reflects on his or her own description and uses imaginative variation or structural description, seeking all possible meanings and divergent perspectives, varying the frames of reference about the phenomenon, and constructing a description of how the phenomenon was experienced (p. 150).
5. The researcher then constructs an overall description of the meaning and the essence of the experience (p. 150).
6. This process is followed first for the researcher's account of the experience and then for that of each participant. After this, a "composite" description is written (p. 150).

Based on the above-presented explanations and their subsequent incorporation in a study on workplace spirituality, the researcher developed the following model (Figure 1), which may serve as an example of a possible phenomenological process with incorporation of interrater reliability as a constructive solidification tool.

Figure 1. Research process in picture.



In the here-discussed phenomenological study, which aimed to establish a broadly acceptable definition of spirituality in the workplace and therefore sought to obtain vital themes that would be applicable in such a work environment, the researcher considered the application of interrater reliability most appropriate at the time when the phenomenological reduction was completed. The meaning clusters also had been formed. Since the most important research findings would be derived from the emergent themes, this seemed to be the most crucial as well as the most applicable part for soliciting interrater reliability. However, the researcher did not submit any pre-classified information to the interraters, but instead provided them the entirety of raw transcribed data with highlights of 3 topical questions from which common themes needed to be derived. In other words, the researcher first performed phenomenological reduction, concluded which questions provided the largest numbers of theme listings, and then submitted the raw version of the answers to these questions to the interraters to find out whether they would come up with a decent amount of similar theme findings. This process will be explained in more detail later in the paper.

Blodgett-McDeavitt (1997) cites one of the prominent researchers in phenomenology, Moustakas, in a presentation of the four main steps of phenomenological processes: epoche, reduction, imaginative variation, and synthesis of composite textural and composite structural descriptions. The way Moustakas' steps can

be considered to correspond with the earlier presented procedure, as formulated by Creswell, is that "epoche" (which is the process of bracketing previous knowledge of the researcher on the topic) happens when the researcher describes his or her own experiences of the phenomenon and thereby symbolically "empties" his or her mind (see Creswell step 1); "reduction" occurs when the researcher finds nonrepetitive, nonoverlapping statements, and groups them into meaning units (Creswell step 2 and 3); "imaginative variation" takes place when the researcher engages in reflection (Creswell step 4); and "synthesis" is applied when the researcher constructs an overall description and formulates his or her own accounts as well as those of the participants (Creswell steps 5 and 6).

Elaborating on the interpretation of epoche, Blodgett-McDeavitt (1997) explains,

> Epoche clears the way for a researcher to comprehend new insights into human experience. A researcher experienced in phenomenological processes becomes able to see data from new, naive perspective from which fuller, richer, more authentic descriptions may be rendered. Bracketing biases is stressed in qualitative research as a whole, but the study of and mastery of epoche informs how the phenomenological researcher engages in life itself. (p. 3)

Although epoche may be considered an effective way for the experienced phenomenologist to empty him or herself and subsequently see the obtained data from a naïve perspective, the chance is that bias is still very present for the less experienced investigator. The inclusion of interrater reliability as a bias reduction tool could therefore lead to significant quality enhancement of the study's findings (as will be discussed below).

Using Interrater Reliability in a Phenomenological Study

Interrater reliability has thus far not been a common application in phenomenological studies. However, once the suggestion was brought up by a team of supervising professors about vital themes in a spiritual workplace, the utilization of this constructive verification tool emerged into an interesting challenge and, at the same time, required a high level of creativeness from the researcher in charge. Because of the uncommonness of using this verification strategy in a qualitative study, especially a phenomenology where the researcher is highly involved in the formulation of the research findings, it was fairly difficult to determine the applicability and positioning of this tool in the study. It was even more complicated to formulate the appropriate approach in calculating this rate, since there were various ways possible for computing it.

The first step for the researcher in this study was to find a workable definition for this verification tool. It was rather obvious that the application of this solidification strategy toward the typical massive amount of descriptive data of a phenomenology would have to differ significantly from the way this tool was generally used in quantitative analysis where kappa coefficients are the common way to go. After in-depth source reviews, the researcher concluded that there was no established consistency to date in defining interrater reliability, since the appropriateness of its outcome depends on

the purpose it is used for. Isaac and Michael (1997) illuminate this by stating that "there are various ways of calculating interrater reliability, and that different levels of determining the reliability coefficient take account of different sources of error" (p. 134). McMillan and Schumacher (2001) elaborate on the inconsistency issue by explaining that researchers often ask how high a correlation should be for it to indicate satisfactory reliability. McMillan and Schumacher conclude that this question is not answered easily. According to them, it depends on the type of instrument (personality questionnaires generally have lower reliability than achievement tests), the purpose of the study (whether it is exploratory research or research that leads to important decisions), and whether groups or individuals are affected by the results (since action affecting individuals requires a higher correlation than action affecting groups).

Aside from the above presented statements about the divergence in opinions with regards to the appropriate correlation coefficient to be used, as well as the proper methods of applying interrater reliability, it is also a fact that most or all of these discussions pertain to the quantitative field. This suggests that there is still intense review and formulation needed in order to determine the applicability of interrater reliability in qualitative analyses, and that every researcher that takes on the challenge of applying this solidification strategy in his or her qualitative study will therefore be a pioneer.

The first step for the researcher of this phenomenological study was attempting to find the appropriate degree of coherence that should exist in the establishment of interrater reliability. It was the intention of the researcher to use a generally agreed upon percentage, if existing, as a guideline in her study. However, after assessing multiple electronic (online) and written sources regarding the application of interrater reliability in various research disciplines, the researcher did not succeed in finding a consistent percentage for use of this solidification strategy. Source included Isaac and Michael's (1997) *Handbook in Research and Evaluation*, Tashakkori and Teddlie's (1998) *Mixed Methodology*, and McMillan and Schumacher's (2001) *Research in Education*; Proquest's extensive article and paper database as well as its digital dissertations file; and other common search engines such as "Google".. Consequently, this researcher presented the following illustrations for the observed basic inconsistency, in applying interrater reliability, as she perceived them throughout a variety of studies, which were not necessarily qualitative in nature.

1. Mott, Etsler, and Drumgold (2003) presented the following reasoning for his interrater reliability findings in their study, *Applying an Analytic Writing Rubric to Children's Hypermedia "Narratives."*

> A comparative approach to the examination of the technical qualities of a pen and paper writing assessment for elementary students' hypermedia-created productsPearson correlations averaged across 10 pairs of raters found acceptable interrater reliability for four of the five subscales. For the four subscales, theme, character, setting, plot and communication, the r values were .59, .55, .49, .50 and .50, respectively (Mott, Etsler, & Drumgold, 2003, ¶1).

2. Butler and Strayer (1998) assert the following in their online-presented research document, administered by Stanford University and titled *The Many Faces of Empathy*.

> Acceptable interrater reliability was established across both dialogues and monologues for all of the verbal behaviors coded. The Pearson correlations for each variable, as rated by two independent raters, are as follows: Average intimacy of disclosure, r =.94, t (8) = 7.79 p < .05; Focused empathy, r =.78, t (14) = 4.66 p < .05; and Shared Affect, r =.85, t (27) = 8.38, p < .05 (¶1).

3. Srebnik, Uehara, Smukler, Russo, Comtois, and Snowden (2002) approach interrater reliability in their study on *Psychometric Properties and Utility of the Problem Severity Summary for Adults with Serious Mental Illness* as follows: "Interrater reliability: A priori, we interpreted the intraclass correlations in the following manner: .60 or greater, strong; .40 to .59, moderate; and less than .40, weak " (¶15).

Through multiple reviews of accepted reliability rates in various studies, this researcher finally concluded that the acceptance rate for interrater reliability varies between 50% and 90%, depending on the considerations mentioned above in the citation of McMillan and Schumacher (1997). The researcher did not succeed in finding a fixed percentage for interrater reliability in general and definitely not for phenomenological research. She contacted the guiding committee of this study to agree upon a usable rate. The researcher found that in the phenomenological studies she reviewed through the Proquest digital dissertation database, interrater reliability had not been applied, although she did find a master's thesis from the Trinity Western University that briefly mentioned the issue of using reliability in a phenomenological study by stating

> Phenomenological research must concern itself with reliability for its results to have applied meaning. Specifically, reliability is concerned with the ability of objective, outside persons to classify meaning units with the appropriate primary themes. A high degree of agreement between two independent judges will indicate a high level of reliability in classifying the categories. Generally, a level of 80 percent agreement indicates an acceptable level of reliability. (Graham, 2001, p. 66)

Graham (2001) then states "the percent agreement between researcher and the student [the external judge] was 78 percent" (p. 67). However, in the explanation afterwards it becomes apparent that this percentage was not obtained by comparing the findings from two independent judges aside from the researcher, but by comparing the findings from the researcher to one external rater. Considering the fact that the researcher in a phenomenological study always ends up with an abundance of themes on his or her list (since he or she manages the entirety of the data, while the external rater only reviews a limited part of the data) calculating a score as high as 78% should not be difficult to obtain depending on the calculation method (as will be demonstrated later in this paper). The citation Graham used as a guideline in his thesis referred to the agreement between

two independent judges and not to the agreement between one independent judge and the researcher.

The researcher of the here-discussed phenomenological study on spirituality in the workplace also learned that the application of this solidification tool in qualitative studies has been a subject of ongoing discussion (without resolution) in recent years, which may explain the lack of information and consistent guidelines currently available.

The guiding committee for this particular research agreed upon an acceptable interrater reliability of two thirds, or 66.7% at the time of the suggestion for applying this solidification tool. The choice for 66.7% was based on the fact that, in this team, there were quantitative as well as qualitative oriented authorities, who after thorough discussion came to the conclusion that there were variable acceptable rates for interrater reliability in use. The team also considered the nature of the study and the multi-interpretability of the themes to be listed and subsequently decided the following: Given the study type and the fact that the interraters would only review part of the data, it should be understood that a correspondence percentage higher than 66.7% between two external raters might be hard to attain. This correspondence percentage becomes even harder to achieve if one considers that there might also be such a high number of themes to be listed, even in the limited data provided, that one rater could list entirely different themes than the other, without necessarily having a different understanding of the text;

The researcher subsequently performed the following measuring procedure:

1. The data gained for the purpose of this study were first transcribed and saved. This was done by obtaining a listing of the vital themes applicable to a spiritual workplace and consisted of interviews taken with a pre-validated interview protocol from 6 participants.
2. Since one of the essential procedures in phenomenology is to find common themes in participants' statements, the transcribed raw data were presented to two pre-identified interraters. The interraters were both university professors and administrators, each with an interest in spirituality in the workplace and, expectedly, with a fairly compatible level of comprehensive ability. These individuals were approached by the researcher and, after their approval for participation, separately visited for an instructional session. During this session, the researcher handed each interrater a form she had developed, in which the interrater could list the themes he found when reviewing the 6 answers to each of the three selected questions. Each interrater was thoroughly instructed with regards to the philosophy behind being an interrater, as well as with regards to the vitality of detecting themes that were common (either through direct wording or interpretative formulation by the 6 participants). The interraters, although acquainted with each other, were not aware of each other's assignment as an interrater. The researcher chose this option to guarantee maximal individual interpretation and eliminate mutual influence. The interraters were thus presented with the request to list all the common themes they could detect from the answers to three particular interview questions. For this procedure, the researcher made sure to select those questions that solicited a listing of words and phrases from the participants. The reason for selecting these questions and their answers was to provide the interraters with a fairly clear and obvious overview of possible themes to choose from.

3. The interraters were asked to list the common themes per highlighted question on a form that the researcher developed for this purpose and enclosed in the data package. Each interrater thus had to produce three lists of common themes: one for each highlighted topical question.

The highlighted questions in each of the six interviews were: (1) What are some words that you consider to be crucial to a spiritual workplace? (2) If a worker was operating at his or her highest level of spiritual awareness, what would he or she actually do? and (3) If an organization is consciously attempting to nurture spirituality in the workplace, what will be present? One reason for selecting these particular responses was that the questions that preceded these answers asked for a listing of words from the interviewees, which could easily be translated into themes. Another important reason was that these were also the questions from which the researcher derived most of the themes she listed. However, the researcher did not share any of the classifications she had developed with the interraters, but had them list their themes individually instead in order to be able to compare their findings with hers.

4. The purpose of having the interraters list these common themes was to distinguish the level of coordinating interpretations between the findings of both interraters, as well as the level of coordinating interpretations between the interraters' findings and those of the researcher. The computation methods that the researcher applied in this study will be explained further in this paper.
5. After the forms were filled out and received from the interraters, the researcher compared their findings to each other and subsequently to her own. Interrater reliability would be established, as recommended by the dissertation committee for this particular study, if at least 66.7% (2/3) agreement was found between interraters and between interraters' and researcher's findings. Since the researcher serves as the main instrument in a phenomenological study, and even more because this researcher first extracted themes from the entire interviews, her list was much more extensive than those of the interraters who only reviewed answers to a selected number of questions. It may therefore not be very surprising that there was 100% agreement between the limited numbers of themes submitted by the interraters and the abundance of themes found by the researcher. In other words, all themes of interrater 1 and all themes of interrater 2 were included in the theme-list of the researcher. It is for this reason that the agreement between the researcher's findings and the interraters' findings was not used as a measuring scale in the determination of the interrater reliability percentage.

A complication occurred when the researcher found that the interraters did not return an equal amount of common themes per question. This could happen because the researcher omitted setting a mandatory amount of themes to be submitted. In other words, the researcher did not set a fixed number of themes for the interraters to come up with, but rather left it up to them to find as many themes they considered vital in the text provided. The reason for refraining from limiting the interraters to a predetermined number of themes was because the researcher feared that a restriction could prompt random choices by each interrater among a possible abundance of available themes, ultimately leading to entirely divergent lists and an unrealistic conclusion of low or no interrater reliability.

To clarify the researcher's considerations a simple example would be if there was a total of 100 obvious themes to choose from and the researcher required the submission of only 15 themes per interrater, there would be no guarantee which part of the 100 available themes each interrater would choose. It could very well be that interrater 1 would select the first 15 themes encountered, while interrater 2 would choose the last 15. If this were the case there would be zero percent interrater reliability, even though the interraters may have actually had a perfect common understanding of the topic. Therefore, the researcher decided to just ask each interrater to list as many common themes he could find among the highlighted statements from the 6 participants. It may also be appropriate to stress here that the researcher explained well in advance to the raters what the purpose of the study was, so there would be no confusion with regards to the understanding of what exactly were considered to be "themes."

Dealing with the problem of establishing interrater reliability with an unequal amount of submissions from the interraters was thus another interesting challenge. Before illustrating how the researcher calculated interrater reliability for this particular case, note the following information:

- Interrater 1 (I1) submitted a total of 13 detected themes for the selected questions.
- Interrater 2 (I2) submitted a total of 17 detected themes for the selected questions.
- The researcher listed a total of 27 detected themes for the selected questions.

Between both interraters there were 10 common themes found. The agreement was determined on two counts: (1) On the basis of exact listing, which was the case with 7 of these 10 themes and (2) on the basis of similar interpretability, such as "giving to others" and "contributing"; "encouraging" and "motivating"; "aesthetically pleasing workplace"; and "beauty" of which the latter was mentioned in the context of a nice environment. The researcher color-coded the themes that corresponded with the two interraters (yellow) and subsequently color-coded the additional themes that she shared with either interrater (green for additional corresponding themes between the researcher and interrater 1 and blue for additional corresponding themes between the researcher and interrater 2). All of the corresponding themes between both interraters (the yellow category) were also on the list of the researcher and therefore also colored yellow on her list.

Before discussing the calculation methods reviewed by this researcher about spirituality in the workplace, it may be useful to clarify that phenomenology is a very divergent and complicated study type, entailing various sub-disciplines and oftentimes described as "the study of essences, including the essence of perception and of consciousness" (Scott, 2002, ¶1). In his presentation of Merleau-Ponty's *Phenomenology of Perception* Scott explains, "phenomenology is a method of describing the nature of our perceptual contact with the world. Phenomenology is concerned with providing a direct description of human experience" (¶1). This may clarify to the reader that the phenomenological researcher is aware that reality is a subjective phenomenon, interpretable in many different ways. Based on this conviction, this researcher did not make any pre-judgments on the quality of the various calculation methods presented below, but merely utilized them on the basis of their perceived applicability to this study type.

The researcher came across various possible methods for calculating interrater reliability described.

Calculation Method 1

Various electronic sources, among which a website from Richmond University (n.d.), mentions the percent agreement between two or more raters as the easy way to calculate interrater reliability. In this case, reliability would be calculated as: (Total # agreements) / (Total # observations) x 100. In the case of this study, the outcome would be: 20/30 x 100 = 66.7%, whereby 20 equals the added number of agreements from both interraters (10 + 10) and 30 equals the added number of observations from both interraters (13 + 17). The recommendation from Posner, Sampson, Ward, and Cheney (1990) is that interrater reliability, R = number of agreements / number of agreements + number of disagreements, also leads to the same outcome. This calculation would be executed as follows: 20 / (20+10) = 2/3 = 66.7%.

Various authors recommend the "confusion matrix," which is a standard classification matrix, as a valid option for calculating interrater reliability. A confusion matrix, according to Hamilton, Gurak, Findlater, and Olive (2003), "contains information about actual and predicted classifications done by a classification system. Performance of such systems is commonly evaluated using the data in the matrix." (¶1) According to these authors, the meaning of the entries in the confusion matrix should be specified as they pertain to the context of the study. In this study the following meanings will be ascribed to the various entries, *a* is the number of agreeing themes that Interrater 1 listed in comparison with Interrater 2; *b* is the number of disagreeing themes that Interrater 1 listed in comparison with Interrater 2; *c* is the number of disagreeing themes that Interrater 2 listed in comparison with Interrater 1; and *d* is the total number of disagreeing themes that both interraters listed.

The confusion matrix that Hamilton et al. (2003) present is similar to the one displayed in Table 1. However, this researcher has specified the entries as recommended by these authors for the purpose of this study.

Table 1

Confusion Matrix 1

		Interrater 1	
		Agree	Disagree
Interrater 2	Agree	a	B
	Disagree	c	D

Hamilton et al. (2003) subsequently present a number of equations relevant to their specific study. The researcher of this study substituted the actual values pertaining to this particular study in the authors' equations and came to some interesting findings:

1. The rate that these authors label as "the accuracy rate" (AC), named this way because it measures the proportion of the total number of findings from Interrater 1 -- the one with the lowest number of themes submitted -- that are "accurate." In this case

"accurate" means in agreement with the submissions of Interrater 2 (adopted from Hamilton et al., 2003, ¶5, and modified toward the values used in this particular study), is calculated as seen below.

AC = (a + d) / (a + b + c + d)
= (10 + 10) / (10 + 3 + 7 + 10)
= 20/30 = 66.7%

2. The rate these authors label as "the true agreement rate:" The title of this rate has been modified by substituting the names of values applicable in this particular study. The true agreement rate was named this way because it measures the proportion of agreed upon themes (10) perceived from the entire number of submitted themes from Interrater 1, the one with the lowest number of submissions (adopted from Hamilton et al., 2003, ¶8, and modified toward the values used in this particular study), is calculated as seen below.

TA = a / (a + b)
= 10 / (10 + 3)
= 10/13 = 76.9%

Dr. Brian Dyre (2003), associate professor of experimental psychology at the University of Idaho also uses the Confusion Matrix for determining interrater reliability. Dyre recommends the following computation under the heading: *Establishing Reliable Measures for Non-Experimental Research.* As mentioned above, this researcher inserted the values that were derived from the interrater reliability test for this particular study about spirituality in the workplace in the recommended columns and rows, presented below as Table 2. The interraters are referred to as R1 and R2.

Table 2

Confusion Matrix 2 with Substitution of Actual Values

		R1		
		Agree	Disagree	
R2	Agree	10	3	13
	Disagree	7	10 (=3+7)	17
		17	13	30

According to Dyre (2003), interrater reliability = (Number of agreeing themes) + (Number of disagreeing themes) / (Total number of observed themes) = (10 + 10) / 30 = 2/3 = 66.7%, which is similar to the earlier discussed accuracy rate (AC) from Hamilton et al. (2003).

Calculation Method 2

Since the interraters did not submit an equal number of observations, as is general practice in interrater reliability measures, the above-calculated rate of 66.7% can be disputed. Although the researcher did not manage to find any written source to base the following computation on, she considered it logical that *in case of unequal submissions, the lowest submitted number of findings from similar data by any of two or more interraters used in a study should be used as the denominator in measuring the level of agreement*. Based on this observation, interrater reliability would be: (Number of common themes) / (Lowest Number of submission) x 100 = 10/13 x 100% = 76.9%.

Rationale for this calculation: if the numbers of submissions by both interraters had varied even more, say 13 for interrater 1 versus 30 for interrater 2, interrater reliability would be impossible to be established even if all the 13 themes submitted by interrater 1 were also on the list of interrater 2. With the calculations as presented under calculation method 1, the outcome would then be: (13 +13) / (30 + 13) = 26/43 = 60.5%, whereby 13 would be the number of agreements and 43 the total number of observations. This does not correspond at all with the logical conclusion that a total level of agreement from one interrater's list onto the other should equal 100%.

If, therefore, the "rational" justification of calculation method 2 is accepted, then interrater reliability is 76.9%, which exceeds the minimally consented rate of 66.7%. Expanding on this reasoning, further comparison leads to the following findings: All 13 listed themes from interrater 1 (13/13 x 100% = 100%) were on the researcher's list and 16 of the 17 themes on interrater 2's list (16/17 X 100% = 94.1%) were also on the list of the researcher. These calculations are based on calculation method 2.

The researcher thought it to be interesting that the percentage of 76.9 between both interraters was also reached in the true agreement rate (TA) as presented earlier by Hamilton et al. (2003).

Calculation Method 3

Elaborating on Hamilton et al.'s (2003) true agreement rate (TA), which is the proportion of corresponding themes identified between both interraters, it is calculated using the equation: TA = a / (a+b), whereby "a" equals the amount of corresponding themes between both interraters and "b" equals the amount of non-corresponding themes as submitted by the interrater with the lowest number of themes. The researcher thought it to be interesting to examine the calculated outcomes in the case that the names of the two interraters would have been placed differently in the confusion matrix. When exchanging the interraters' places in the matrix the outcome of this rate turned out to be different, since the value substituted for "b" now became that of the number of non-corresponding themes, as submitted by the interrater with the highest number of themes. In fact, the new computation led to an unfavorable, but also unrealistic interrater reliability rate of 58.8%. The "unrealistic" reference lies in the fact that it becomes apparent that the interrater reliability rate, in the case of the above-mentioned substitution, starts turning out extremely low as the submission numbers of the two interraters start differing to an increasing degree. In such a case, it does not even matter anymore whether the two interraters have full correspondence as far as the submissions

of the lowest submitter goes: The percentage of the interrater reliability, which is supposed to reflect the common understanding of both interraters, will decrease to almost zero.

To illustrate this assertion, the confusion matrix is presented in Table 3 with the names of the interraters switched.

Table 3

Confusion Matrix with Names of Interraters Switched

		Interrater 2	
		Agree	Disagree
Interrater 1	Agree	a	b
	Disagree	c	d

With this exchange, the outcome for TA changes significantly:

1. The rate that these authors label as "the accuracy rate" (AC), remains the same:
   AC = (a + d) / (a + b + c + d)
   = (10 + 10) / (10 + 3 + 7 + 10)
   = 20/30 = 66.7%

2. "The true agreement rate" (title name substituted with names of values applicable in this study), is calculated as follows.
   TC = a / (a + b)
   = 10 / (10 + 7)
   = 10/17 = 58.8%

In this study, TA rationally presented a rate of 76.9%, which was higher than the minimum requirement of 66.7% in both, calculation methods 1 and 2. On the other hand it is demonstrated in the new true agreement rate here that the less logical process of exchanging the interraters' positions to where the *highest* number of submissions would be used as the common denominator instead of the *lowest* (see first part of calculation method 3), delivered a percentage below the minimum requirement. As a reminder to the reader the irrationality of using the highest number of submissions as the denominator may serve the example given under the "rationale" section for calculation method 2, in which numbers of submissions would diverge significantly (30 vs. 13). It is the researcher's opinion that this new suggested "computation of moderation" would lead to the following outcome for the true agreement reliability rate (TAR):

TAR = ((TA-1) + (TA-2)) / 2
= (76.9% + 58.8%) / 2
= 135.7% / 2 = 67.9%

It was the researcher's conclusion that whether the reader considers calculation method 1, calculation method 2, or calculation method 3 as the most appropriate one for this particular study, all three methods demonstrated that there was sufficient common

understanding and interpretation of the essence of the interviewees' declarations, as they all resulted in outcomes equal to, or greater than, 66.7%. Hence, for this study interrater reliability could be considered established.

Recommendations

1. The researcher of this study has found that although interraters in a phenomenological study, and presumably generally in qualitative studies, can very well select themes with a similar understanding of essentials in the data she also found that there are three major attention points to address in order to enhance the success rate and swiftness of the process:1. The data to be reviewed by the interraters should be only a segment of the total amount, since data in qualitative studies are usually rather substantial and interraters usually only have limited time.2. The researcher will need to understand that there are different configurations possible in the packaging of the themes as listed by the various interraters, so that he or she will need to review the context in which these themes are listed in order to determine their correspondence (Armstrong et al., 1997). In this paper the researcher gave examples of themes that could be considered similar, although they were "packaged" different by the interraters, such as "giving to others" and "contributing;" "encouraging" and "motivating;" "aesthetically pleasing workplace;" and "beauty," of which the latter was mentioned in the context of a nice environment.
2. In order to obtain results with similar depth from all raters, the researcher should set standards in the number of observations to be listed by the interraters as well as the time allotted to them. The fact that these confines were not specified to the interraters resulted in a diverged level of input: One interrater spent only two days in listing the words and came up with a total of 13 themes and the other interrater spent approximately one week in preparing his list and consequently came up with a more detailed list of 17 themes. Although there was a majority of congruent themes between the two interraters (there were 10 common themes between both lists), the calculation of interrater reliability was complicated by the unequal numbers of submissions. All interrater reliability calculation methods assume equal numbers of submissions by the interraters. The officially recognized reliability rate of 66.7% for this study is therefore lower than it would have been when both interraters had been limited to a pre-specified number of themes to be listed. If, for example, both interraters had been required to select 15 themes within an equal time span of, say, one week, the puzzle regarding the use of either the lowest or highest common denominator would be resolved because there would be only one denominator, as well as an equal level of input from both interraters. If, in this case, the interraters came up with 12 common themes out of 15, the interrater reliability rate could be easily calculated as 12/15 = .8 = 80%. Even in the case of only 10 common themes on a total required submission of 15, the rate would still meet the minimum requirements: 10/15 = .67 = 66.7%. This may be valuable advice for future applications of this valuable tool to qualitative studies.4. The solicited number of submissions from the interraters should be set as high as possible, especially if there is a multiplicity of themes to choose from. If the solicited number is kept too low it may be that two raters have perfectly similar understanding of the text yet submit

different themes, which may erroneously elicit the idea that there was not enough coherence in the raters' perceptions and, thus, no sufficient interrater reliability.

3. The interraters should have at least a reasonable degree of similarity in intelligence, background, and interest level in the topic in order to ensure a decent degree of interpretative coherence. It would further be advisable to attune the educational and interest level of the interraters to the target group of the study, so that the reader could encounter a greater level of recognition with the study topic as well as the findings.

Conclusion

As mentioned previously, interrater reliability is not a commonly used tool in phenomenological studies. Of the eight phenomenology dissertations that this researcher reviewed, prior to embarking on her own experiential journey, none applied this instrument of control and solidification. This was possibly attributable to the fact that various qualitative oriented scholars have asserted in the past years that it is difficult to obtain consistency in qualitative data analysis and interpretation (Armstrong et al., 1997). These scholars instead, introduced a variety of "new criteria for determining reliability and validity, and hence ensuring rigor, in qualitative inquiry" (Morse et al., 2002, p. 2). Unfortunately, the majority of these criteria are either of a "post hoc" (evaluative) nature, which entails that they are applied after the study had been executed and correction is not possible anymore; or of a non-rigorous nature, such as member checks, which are merely used as a confirmation tool for the study participants regarding the authenticity of the provided raw data, but have nothing to do with the data analysis procedures (Morse et al.). However, having been confronted by the guiding committee in a phenomenological study on spirituality in the workplace, with the application of this tool as an enhancement of the reliability of the findings as well as a bias reduction mechanism, the researcher found that the establishment of interrater reliability or interrater agreement was a major solidification of the themes that were ultimately listed as the most significant ones in this study.

It is the researcher's opinion that the process of interrater reliability should be applied more often to phenomenological studies, in order to provide them with a more scientifically recognizable basis. Up to now, it is still a general perception that qualitative study, a category to which phenomenology belongs, is less scientifically grounded than quantitative study. This perception is supported by the arguments from various scholars that different reviewers cannot coherently analyze a single package of qualitative data. However, the researcher of this particular study has found that the interraters, given the prerequisite of a certain minimal similarity in educational and cultural background as well as interest, could very well select themes with a similar understanding of essentials in the data. This conclusion is shared with Armstrong et al. (1997), who came to similar findings in an empirical study in which they attempted to detect the level to which various external raters could detect themes from similar data and demonstrate similar interpretations. The two main prerequisites presented by Armstrong et al., entailing data limitation and contextual interpretability, were similar to those from the researcher in this phenomenological study. These prerequisites were presented in this paper in the recommendations section.

An interesting lesson from this experience for the researcher was that the number of observations to be listed by the interraters, as well as the time allotted to the interraters, should preferably be kept synchronous. Yet, one might attempt to set as high a number of submissions as possible, due to the risk of too widely varied choices to be selected by interraters, if there are many themes available. This may happen in spite of perfect common understanding between interraters and may, henceforth, wrongfully educe the idea that there is not enough consistency in comprehension between the raters and, thus, no interrater reliability. The justifications for this argument are also presented in the recommendations section of this paper.

References

Armstrong, D., Gosling, A., Weinman, J., & Martaeu, T. (1997). The place of inter-rater reliability in qualitative research: An empirical study. *Sociology, 31*(3), 597-606.

Association for Spirit at Work (2005). *The professional association for people involved with spirituality in the workplace*. Retrieved February 20, 2005, from http://www.spiritatwork.com/aboutSAW/profile_JudiNeal.htm

Blodgett-McDeavitt, C. (1997, October). *Meaning of participating in technology training: A phenomenology*. Paper presented at the meeting of the Midwest Research-to-Practice Conference in Adult, Continuing and Community Education, Michigan State University, East Lansing, MI. Retrieved January 25, 2003, from http://www.iupui.edu/~adulted/mwr2p/prior/blodgett.htm

Butler, E. A., & Strayer, J. (1998). *The many faces of empathy*. Poster presented at the annual meeting of the Canadian Psychological Association, Edmonton, Alberta, Canada.

Colorado State University. (1997). *Interrater reliability*. Retrieved April 8, 2003, from http://writing.colostate.edu/guides/research/relval/com2a5.cfm

Creswell, J. (1998). *Qualitative inquiry and research design: Choosing among five traditions*. Thousand Oaks, CA: Sage.

Dyre, B. (2003, May 6). *Dr. Brian Dyre's pages*. Retrieved November 12, 2003, from http://129.101.156.107/brian/218%20Lecture%20Slides/L10%20research%20designs.pdf

A phenomenological study of quest-oriented religion. Retrieved September 5, 2004, from http://www.twu.ca/cpsy/Documents/Theses/Matt%20Thesis.pdf

Hamilton, H., Gurak, E., Findlater, L., & Olive, W. (2003, February 7). *The confusion matrix*. Retrieved November 16, 2003, from http://www2.cs.uregina.ca/~hamilton/courses/831/notes/confusion_matrix/confusion_matrix.html

Isaac, S., & Michael, W. (1997). *Handbook in research and evaluation* (Vol. 3). San Diego, CA: Edits.

McMillan, J., & Schumacher, S. (2001). *Research in education* (5th ed.). New York: Longman.

Ian I. Mitroff. (2005). Retrieved February 20, 2005, from the University of Southern California Marshall School of Business web site: http://www.marshall.usc.edu/web/MOR.cfm?doc_id=3055

Morse, J. M., Barrett, M., Mayan, M., Olson, K., & Spiers, J. (2002). Verification strategies for establishing reliability and validity in qualitative research. *International Journal of Qualitative Methods, 1*(2), 1-19.

Mott, M. S., Etsler, C., & Drumgold, D. (2003). Applying an analytic writing rubric to children's hypermedia "narratives". *Early Childhood Research & Practice,5*(1*)* Retrieved September 25, 2003, from http://ecrp.uiuc.edu/v5n1/mott.html

Myers, M. (2000, March). Qualitative research and the generalizability question: Standing firm with proteus. *The Qualitative Report, 4*(3/4), Retrieved March 10, 2005, from http://www.nova.edu/ssss/QR/QR4-3/myers.html

Posner, K. L., Sampson, P. D., Ward, R. J., & Cheney, F. W. (1990, September). *Measuring interrater reliability among multiple raters: An example of methods for nominal data.* Retrieved November 13, 2003, from http://schatz.sju.edu/multivar/reliab/interrater.html

Richmond University. (n.d.). *Interrater reliability*. Retrieved November 13, 2003, from http://www.richmond.edu/~pli/psy200_old/measure/interrater.html

School of Business at the University of New Haven. (2005). *Judi Neal Associate Professor*. Retrieved February 20, 2005, from http://www.newhaven.edu/faculty/neal/

Scott, A. (2002). *Merleau-Ponty's phenomenology of perception*. Retrieved September 5, 2004, from http://www.angelfire.com/md2/timewarp/merleauponty.html

Srebnik, D. S., Uehara, E., Smukler, M., Russo, J. E., Comtois, K. A., & Snowden, M. (2002, August). Psychometric properties and utility of the problem severity summary for adults with serious mental illness. *Psychiatric Services 53,* 1010-1017. Retrieved March 4, 2005, from http://ps.psychiatryonline.org/cgi/content/full/53/8/1010

Tashakkori, A., & Teddlie, C. (1998). *Mixed methodology* (Vol. 46). Thousand Oaks, CA: Sage.

Van Manen, M. (2002a). *Phenomenological inquiry.* Retrieved September 4, 2004, from http://www.phenomenologyonline.com/inquiry/1.html

Van Manen, M. (2002b). *Sources of meaning.* Retrieved September 4, 2004, from http://www.phenomenologyonline.com/inquiry/49.html

Appendix A

Interview Protocol

Project: Spirituality in the Workplace: Establishing a Broadly Acceptable Definition of this Phenomenon

Time of interview:
Date:
Place:
Interviewer:
Interviewee:
Position of interviewee:

To the interviewee:

Thank you for participating in this study and for committing your time and effort. I value the unique perspective and contribution that you will make to this study.

My study aims to establish a broadly acceptable definition of "spirituality in the workplace" by exploring the experiences and perceptions of a small group of recognized interviewees, who have had significant exposure to the phenomenon, either through practical or theoretical experience. You are one of these icons identified. You will be asked for your personal definitions and perceived essentials (meanings, thoughts, and backgrounds) regarding spirituality in the workplace. I am looking for accurate and comprehensive portrayals of what these essentials are like for you: your thoughts, feelings, insights, and recollections that might illustrate your statements. Your participation will hopefully help me understand the essential elements of "spirituality in the workplace."

Questions

1. Definition of Spirituality in the Workplace

1.1 How would you describe spirituality in the workplace?
1.2 What are some words that you consider to be crucial to a spiritual workplace?
1.3 Do you consider these words applicable to all work environments that meet your personal standards of a spiritual workplace?
1.4 What is essential for the experience of a spiritual workplace?

2. Possible structural meanings of experiencing spirituality in the workplace?

2.1 If a worker was operating at his or her highest level of spiritual awareness, what would he or she actually do?
2.2 If a worker was operating at his or her highest level of spiritual awareness, what would he or she <u>not</u> do?
2.3 What is easy about living in alignment with spiritual values in the workplace?
2.4 What is difficult about living in alignment with spiritual values in the workplace?

3. Underlying themes and contexts for the experience of a spiritual workplace

3.1 If an organization is consciously attempting to nurture spirituality in the workplace, what will be present?
3.2 If an organization is consciously attempting to nurture spirituality in the workplace, what will be absent?

4. General structures that precipitate feelings and thoughts about the experience of spirituality in the workplace.

4.1 What are some of the *organizational* reasons that could influence the transformation from a workplace that does <u>not</u> consciously attempt to nurture spirituality and the human spirit to one that does?
4.2 From the *employee's perspective*, what are some of the reasons to transform from a worker who does <u>not</u> attempt to live and work with spiritual values and practices to one that does?

5. Conclusion

Would you like to add, modify or delete anything significant from the interview that would give a better or fuller understanding toward the establishment of a broadly acceptable definition of "spirituality in the workplace"

Thank you very much for your participation.

Author Note

Joan Marques was born in Suriname, South America, where she made a career in advertising, public relations, and program hosting. She founded and managed an advertising and P.R. company as well as a foundation for women's awareness issues. In 1998 she immigrated to California and embarked upon a journey of continuing education and inspiration. She holds a Bachelors degree in Business Economics from M.O.C. in Suriname, a Master's degree in Business Administration from Woodbury University, and a Doctorate in Organizational Leadership from Pepperdine University. Her recently completed dissertation was centered on the topic of "spirituality in the workplace." Dr. Marques is currently affiliated to Woodbury University as an instructor of Business & Management. She has authored a wide variety of articles pertaining to workplace contentment for audiences in different continents of the globe. Joan Marques, 712 Elliot Drive # B, Burbank, CA 91504; E-mail: jmarques01@earthlink.net; Telephone: (818) 845 3063

Chester H. McCall, Jr., Ph.D. entered Pepperdine University after 20 years of consulting experience in such fields as education, health care, and urban transportation. He has served as a consultant to the Research Division of the National Education Association, several school districts, and several emergency health care programs, providing survey research, systems evaluation, and analysis expertise. He is the author of two introductory texts in statistics, more than 25 articles, and has served on the faculty of The George Washington University. At Pepperdine, he teaches courses in data analysis, research methods, and a comprehensive exam seminar, and also serves as chair for numerous dissertations. Email: cmccall@pepperdine.edu

Article Citation

Marques, J. F. (2005). The application of interrater reliability as a solidification instrument in a phenomenological study. *The Qualitative Report 10*(3), 439-462. Retrieved [Insert date], from http://www.nova.edu/ssss/QR/QR10-4/marques.pdf

---