E/B/R ECONOMIC AND BUSINESS REVIEW

Volume 14 | Issue 1 Article 3

4-24-2012

A perspective on scale development in entrepreneurship research

Alenka Slavec

Mateja Drnovšek

Recommended Citation

Slavec, A., & Drnovšek, M. (2012). A perspective on scale development in entrepreneurship research. *Economic and Business Review, 14*(1). https://doi.org/10.15458/2335-4216.1203

A PERSPECTIVE ON SCALE DEVELOPMENT IN ENTREPRENEURSHIP RESEARCH

ALENKA SLAVEC[1]
MATEJA DRNOVŠEK[2]

ABSTRACT: *To develop a measurement scale that would lead to valid and reliable results is a tasking goal in any research field. We build from theoretical findings that are advanced in the measurement and scale development theory to review the type of measures used and scale development procedures of measures reported in top-tier entrepreneurship journals during the years 2009 and 2010. The in-depth review showed that not all steps in scale development in entrepreneurship research are always accomplished although every step is crucial when aiming for a sound measure. Additionally, the study showed that all too often measures, that have not been validated, are being used. Thus, this review serves as a reference for researchers to apply methodologically rigorous procedures for scale development in entrepreneurship research and to use previously validated scales. With this in mind, we propose hands on guidelines for scale development in entrepreneurship research in the Appendix of the paper.*

Key words: *scale development, measurement procedure, reliability, validity, entrepreneurship, construct*
JEL classification: L26

1. INTRODUCTION

To develop a measurement scale that would lead to valid and reliable results is a challenging task in any research field. It takes time and reflection. Several scholars argue that effective measurement is a cornerstone of scientific research (e.g. DeVellis, 2003; Netemeyer, Bearden, & Sharma, 2003) and is a central component of good assessment of latent variables (C. R. Reynolds, 2010) and empirical research (Crook, Shook, Madden, & Morris, 2009). Besides that, reliable and valid measures contribute to the legitimacy and development of a research field. Additionally, C. R. Reynolds (2010) stated that the ability and skill to measure variables accurately is a cornerstone to progress in science. Nevertheless, empirical articles that use rigorous methodological procedures, besides being firmly grounded in theory, receive more citations and are more important to the development of science (Colquitt & Zapata-Phelan, 2007). The field of entrepreneurship is no exception (Kuskova, Podsakoff, & Podsakoff, 2011).

The progress of the measurement issue in the scientific discipline of entrepreneurship has been investigated by several scholars (e.g. Chandler & Lyon, 2001; Crook et al.,

[1] University of Ljubljana, Faculty of Economics, Slovenia, Ljubljana, e-mail alenka.slavec@ef.uni-lj.si
[2] University of Ljubljana, Faculty of Economics, Slovenia, Ljubljana, e-mail mateja.drnovsek@ef.uni-lj.si

2009; Crook, Shook, Morris, & Madden, 2010; Dean, Shook, & Payne, 2007; Kuskova et al., 2011; Low, 2001; Low & MacMillan, 1988). They found that as a young and rapidly growing field (Dean et al., 2007; Shane & Venkataraman, 2000) entrepreneurship has evidenced improvements in research design and constructs measurement (Crook et al., 2010). Despite the methodological progress, we are not there yet (Crook et al., 2010). The main criticism is that the field has not undertaken empirical research with the same rigor as other fields (Low, 2001) and extant research relies far too heavily on measures that do not allow for or report on the assessment of reliability, validity, model fit, measurement error (Crook et al., 2009; Crook et al., 2010) and other important issues in construct development like content domain specification, item pool generation, and pilot study conduction. Still, the appropriate use of and sufficient sophistication in the use of analytical techniques is a critical component to the advancement and legitimacy of the entrepreneurship field (Dean et al., 2007). While research design and construct measurement in entrepreneurship research have already been evaluated, a study that would assess scale development approaches in the scale development process for the population of entrepreneurship papers has not been undertaken yet.

With this study we aim to contribute towards more rigorous scale development in entrepreneurship research by investigating crucial steps in the scale development procedure and pointing out specific problems associated with each step. Additionally, in the Appendix of the paper we provide hands on guidelines for developing new measures for researchers to get an overview of the crucial steps that should be undertaken when developing new measures. Moreover, we also aim at spurring the awareness that validated scales should be used when testing constructs. The motivation for this research springs from authors' systematic review of studies published and observation that in entrepreneurship journals many times measures have not been properly conceptualized, operationalized or applied. When proposing a new measure some important steps in the scaling procedure have been overlooked. Of concern is also the fact that all too often not-validated measures are extensively used. Both represent a problem since researchers in subsequent studies may use these measures and results may be questionable. Although we try to critically evaluate scaling procedures reported in papers, we note our debt to previous researchers who performed the exploratory work upon which more recent research is built (Chandler & Lyon, 2001; Low, 2001; Low & MacMillan, 1988). By assessing the methodological part of previous works we aim to identify critical deficiencies in scaling procedures and propose alternatives to overcome them with guidelines in the form of a ten-step methodology that we attach in the Appendix of this paper. We partially adopt Low and MacMillan`s (1988), Chandler and Lyon`s (2001), and Crook et al.`s (2010) research setting and evaluate contributions and shortcomings of the methodological part of the research design specification. In contrast to their lens of view, we focus on the type of measures applied and the steps in scale development reported in the reviewed papers. These findings provide an admonishment for future research and in particular for scale development in entrepreneurship research.

In what follows, we discuss and evaluate the measures used and research setting of scale development applied in entrepreneurship research. Our database includes a sam-

ple of 197 papers published in 2 top-tier entrepreneurship journals (*Journal of Business Venturing* and *Entrepreneurship Theory and Practice*) from January 2009 to December 2010. In so doing we contribute to the entrepreneurship discipline by evaluating the current state of the type of measures used and scale development procedures applied in entrepreneurship research and by identifying problematic issues in the measurement part of entrepreneurship papers. To overcome the problematic issues we provide an overview of the crucial steps in scale development in the form of a ten-step three-phased methodology in the Appendix of the paper and offer implications for future research.

2. REVIEW OF MEASURES IN ENTREPRENEURSHIP JOURNALS

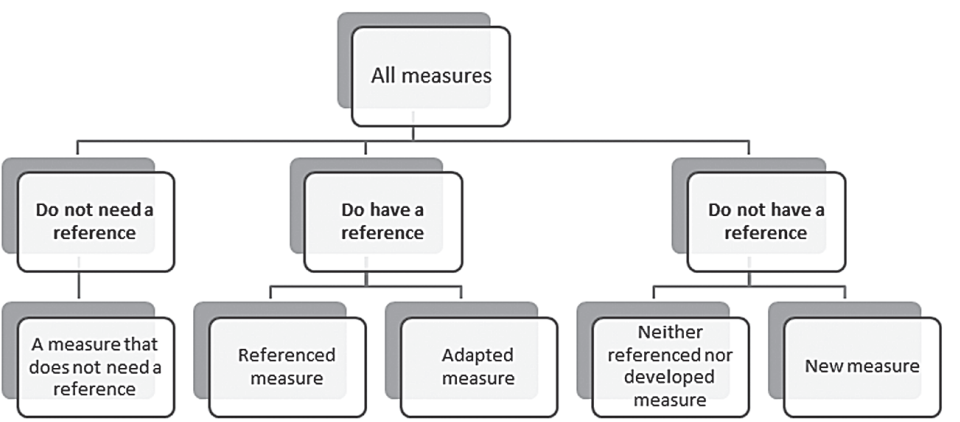
In our study we partly followed Low and MacMillan`s (1988), Chandler and Lyon`s (2001), and Crook et al.`s (2010) framework of reviewing papers in entrepreneurship journals. The starting point of our study was to take inventory of scale development procedures and type of measures reported. Specifically, we addressed the following two questions. (1) Are measures that are used verified, i.e. referenced? (2) If not, what was the procedure for developing new measures? For this analysis we included papers published from January 2009 to December 2010 in the following two journals: *Journal of Business Venturing* and *Entrepreneurship Theory & Practice*. The selection of journals was based on the ranking scale of SSCI journals in entrepreneurship research as on October 2010; we analyzed the first two ranked entrepreneurship journals on this scale. The selection of journals is consistent with other research settings in entrepreneurship (e.g. Crook et al., 2009; Crook et al., 2010; Dean et al., 2007). In *Journal of Business Venturing* we reviewed Volume 24, Issue 1-6 and Volume 25, Issue 1-6 and in *Entrepreneurship Theory & Practice* we reviewed Volume 33, Issue 1-6 and Volume 34, Issue 1-6. In 2009 and 2010 a total of 197 papers were published in the two reviewed journals. The detailed results on the number of empirical papers and measures reviewed by journal and year are presented in Table 1 in section 2.2 Results. This time span seems reasonable to get an appropriate review of the current state of the measurement issue in entrepreneurship research.

2.1 Method

The content analysis of the 197 papers was conducted as follows. First, each paper was categorized as either empirical or conceptual. For a paper to be categorized as empirical it had to include some kind of data or data analysis (Chandler & Lyon, 2001). A category of "conceptual paper" was assigned to those papers that included literature reviews, untested theoretical models, mathematical models (Chandler & Lyon, 2001), conceptual papers, case studies, teaching cases and review papers. Studies that both presented and tested theory with empirical data were categorized as empirical studies (Chandler & Lyon, 2001). When a paper was categorized as conceptual, the review ended at this stage while empirical papers were further assessed.

Empirical papers were analyzed using a rating form for measures developed specifically for this study. Measures were coded as dummy (0,1) variables; we either found evidence supporting the categorization (1) or we did not (0) (Chandler & Lyon, 2001). For each measure we analyzed whether it was correctly referenced (1) or not (0). A *referenced measure* is a measure that has been previously used by other researchers and studies. Further, some measures were categorized as *adapted measures*, i.e. the ones that have a reference and have been slightly modified from its original form and authors reported on that. In the rating form in the appropriate column an adapted measure was noted down as a 1; if this was not the case we wrote a 0. For those measures that did not have a reference, we analyzed whether they were developed throughout the paper and two additional types of measures were identified at this stage. The first is a *neither referenced nor developed measure*. This is a measure that neither has references of previous uses (i.e. does not report on previous uses of the measure) nor is developed throughout the study (i.e. does not report on scale development procedures). If a measure was qualified as a *neither referenced nor developed measure* it got a 1, if this was not the case we wrote a 0. The second is a *new measure* which is a measure that is developed in the study, reports on some or all steps in scale development procedures and has not been previously used. Again, if a measure was categorized as a *new measure*, we noted down a 1, otherwise it got a 0. Finally, some measures were categorized as *measures that do not need a reference*. A measure of this type is qualified as the one that does not need to be validated or to be referenced. Examples for such measures are: gender, age, education, marital status, race, industry classification, parental entrepreneurship, firm legal form, etc. A single measure was categorized only as one type of measure (e.g. a *referenced measure*) and could not be categorized as several types of measures (e.g. not a *referenced measure* and a *new measure* at the same time). Figure 1 summarizes the categorization of the measures, whereas Table 2 in section 2.2 Results reports the results on the categorization of measures.

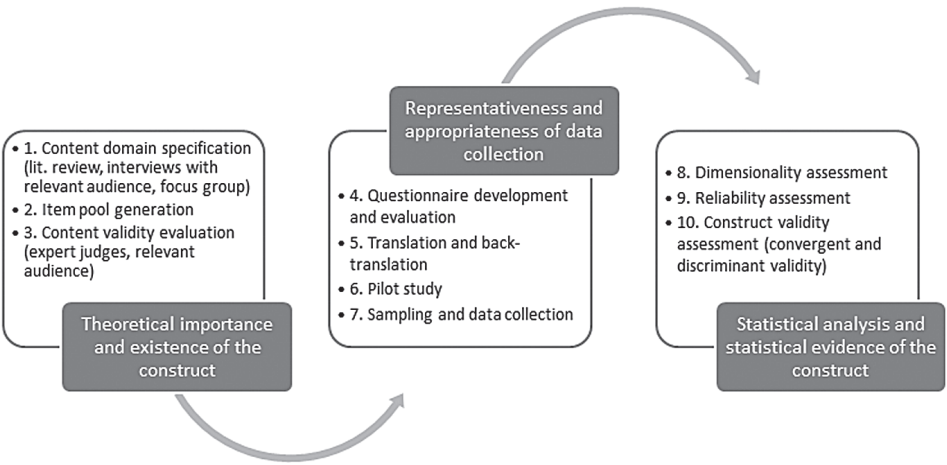
Figure 1: *Categorization of measures*



For measures that were categorized as *new measures* we investigated the scale development procedure that was reported in papers. Based on the extant literature regarding

scale development (e.g. Bagozzi & Edwards, 1998; Carmines & Zeller, 1979; Churchill, 1979; DeVellis, 2003; Hinkin, 1998; Netemeyer et al., 2003; Nunnally & Bernstein, 1994; Pedhazur & Pedhazur Schmelkin, 1991) we identified ten steps that were crucial for developing new measures for entrepreneurship-related studies. If a step was reported in the paper, we denoted it in the rating form with a 1. We noted down a 0 if the step has not been reported. The ten steps were grouped into three phases and are presented in Figure 2.

Figure 2: *Ten steps and three phases in scale development*



The ten scale development steps are grouped into three phases. The first phase regards the theoretical importance and existence of the construct, the second phase deals with the representativeness and appropriateness of data collection, and the third phase regards the statistical analysis and statistical evidence of the construct. The ten steps that comprise the three phases were the following (for more detailed descriptions of the ten steps see the Appendix). The first step regarded the *content domain specification* since the development of a new measure starts with outlining the domain of the new construct. It is achieved by an in-depth literature review (Netemeyer et al., 2003). This step consists of stating what the researcher means by the use of a particular word by providing a definition of the new construct (Nunnally & Bernstein, 1994). If the domain of the new construct was specified, we noted down a 1 in the rating form. If this was not the case, a 0 was marked. In the second step a researcher generates a *pool of potential items* that sample the domain of the construct. From this pool of items the new scale will be derived. In the third step we investigated whether the *content validity* has been evaluated. Content validity refers to the adequacy of sampling the domain of the construct (Nunnally & Bernstein, 1994), i.e. the assessment of the adequacy of the proposed items from the item pool by the relevant audience. The first three steps represent the first phase of scale development, i.e. the theoretical importance and existence of the construct.

The second phase on the representativeness and appropriateness of data collection is comprised by the next four steps (steps from 4 to 7). The fourth and fifth steps are optional since they are dependent on the type of research setting. If the data collecting method was a survey, the forth step investigated whether the papers reported on *questionnaire development and evaluation* (1) or not (0). The fifth step is mandatory for cross-national studies where different languages are spoken. Here, the translation and back-translation is required and we investigated whether papers reported on *translation and back-translation* (1) or not (0).

The sixth step investigated whether a *pilot study* was performed (1) or not (0). This step is recommendable and useful to test the proposed questionnaire and measure. In the pilot study a researcher identifies potential problems with the questionnaire (Dillman, Smyth, & Christian, 2009) and obtains a pre-result of the reliability of the new measure and correlations among items. In the seventh step we analyzed whether *sampling and data collection* was reported (1) or not (0). A representative sample is crucial for the quality of results. Not much less important is the procedure of collecting data.

In the final three steps which comprise the third phase on the statistical analysis and statistical evidence of the construct, we investigated whether *dimensionality* (1), *reliability* (1), and *construct validity* (1) were assessed or not (0). A measure`s dimensionality is concerned with the homogeneity of items (Netemeyer et al., 2003). Reliability concerns the extent to which a measurement procedure yields the same results on repeated trials (Carmines & Zeller, 1979). *Construct validity* is a scale`s ability to measure what it is supposed to measure (Haynes, Richard, & Kubany, 1995; Hinkin, 1998). Despite some disagreements about the classification of different types of validity, two main types of construct validity should be evaluated – convergent and discriminant validity. *Convergent validity* refers to the degree to which different measures of the same construct are correlated (Netemeyer et al., 2003). *Discriminant validity* assesses the degree to which two measures design to measure similar, but conceptually different constructs, are related (Netemeyer et al., 2003). If papers reported on other types of validity these additional validities were noted down.

2.2 Results

In Volume 24 and Volume 25 in *Journal of Business Venturing* and in Volume 33 and Volume 34 in *Entrepreneurship Theory & Practice* a total of 197 papers were published. Among 197 papers, 98 papers were classified as empirical and 99 as conceptual. In the 98 empirical papers 1171 measures were reported. The number of empirical papers and measures reviewed per journal per year is shown in Table 1.

Table 1: *Number of empirical papers and measures reviewed by journal and year*

Journal/Year	2009		2010		Total	
	# of papers	# of measures	# of papers	# of measures	# of papers	# of measures
Journal of Business Venturing	21	205	25	399	46	604
Entrepreneurship Theory & Practice	31	328	21	239	52	567
Total	**52**	**533**	**46**	**638**	**98**	**1171**

Most of the empirical papers reported that several measures have been used in their studies and a total of 1171 measures have been identified. From the 1171 measures, 370 (31.6%) measures were categorized as *measures that do not need a reference*. The total number of correctly *referenced measures* was 299 (25.5%), while the total number of *adapted measures* was 42 (3.6%). The most critical category of measures is the one that has measures that *neither have a reference nor are developed* within the study. A total of 383 (32.7%) measures were of this kind. Finally, 77 (6.6%) measures were *new measures* that have been developed for the purpose of the study. The summary of the results is presented in Table 2.

Table 2: *Number and percentage of categorized measures*

Type of measure	Total number of measures	Percentage of measures (in %)
A measure that does not need a reference	370	31.6
Referenced measure	299	25.5
Adapted measure	42	3.6
Neither referenced nor developed measure	383	32.7
New measure	77	6.6
Total	**1171**	**100.0**

Most of the measures that were categorized as *measures that do not need a reference* were control variables and/or with obvious structures, e.g. gender, age in number of years, highest level of education achieved, industry with SIC classification, whether a business is a family business or not, etc. Some *adapted measures* reported on reliability and some also on other steps in scale development, but these measures did not categorize as newly developed measures. Our analysis revealed that almost a third of measures (32.7%) *neither had a reference of previous uses nor were developed* for the purpose of their research. These measures are regarded as problematic since they cannot be considered reliable, valid and effective. There are three main reasons for this thesis. The first is that if no previous research used these measures, we cannot know if they are good measures and confidence in the results of such studies is limited (Crook et al., 2010). While we are aware that researchers might have used referenced measures (and not *neither referenced nor developed measure*) or that steps in scale development procedure have been undertaken, we were unable to control for these shortcomings if papers did not report on that. Secondly, these measures also did not go through steps in scale development which are

essential for the evaluation of the effectiveness of a measure. Finally, subsequent studies may use these measures and results might be biased.

The 77 *new measures* were evaluated in accordance to the ten steps in scale development. The results of the study show that a large proportion of new measures overlooked some steps in scale development. The summary of the number and percentage of measures that accomplished each of the ten steps is presented in Table 3. In what follows we describe these results in details.

Table 3: *Summary of scale development steps reported in papers that developed new measures*

Step in scale development	Number of measures that adopted specific steps in scale development	Percentage of measures (in %) that adopted specific steps in scale development
(1) content domain specification	13	16.9
(2) item pool generation	19	24.7
(3) content validity evaluation	13	16.9
(4) questionnaire development and evaluation	13	16.9
(5) translation and back-translation of the questionnaire	11	14.3
(6) pilot study performance	8	10.4
(7) sampling and data collection	13	16.9
(8) dimensionality assessment	20	26.0
(9) reliability assessment	70	90.9
(10) construct validity assessment	28	36.4

Most of the papers reported the reliability assessment of the new measure (90.9%). Construct validity was assessed in 36.4% of cases. 26.0% of measures have been evaluated for their dimensionality, while item pool generation has been reported in 24.7% of cases. Other steps have been accomplished in less than 20% of cases.

While scale development begins with the specification of the domain, i.e. what is the construct about and what is not to be included in the construct, and with proposing a definition of the new construct (Nunnally & Bernstein, 1994), only 16.9% of measures have accomplished the first step. It is important to note that a clearly specified content domain is the basis for all subsequent steps (Netemeyer et al., 2003). Moreover, a lot of measures have omitted the item pool generation step. These studies reported only which items have been used although these items may not be the most appropriate ones. Accordingly, C. R. Reynolds (2010) emphasizes that only proposing a limited number of items and beginning research on a new construct is not appropriate. In this way a researcher ignores crucial initial steps in scale development (C. R. Reynolds, 2010). Of great concern is the fact that only 16.9% of measures have been evaluated for their content validity, i.e. items have been administered to the relevant audience like entrepreneurship scholars, entrepreneurs, entrepreneurship students, experts from entrepreneurship and psychometric specialists to test if the domain has been sampled correctly and if instructions and

formats are appropriate and clear. While at this stage suggestions by judges are made for additional inclusion of items or for their eventual exclusion and modification, most of the papers did not report on that.

Step 4 investigated whether papers reported on questionnaire development, evaluation and refinement. While this is not a mandatory step for studies that collect data in other ways, 16.9% of new measures have reported on the questionnaire issue. Additionally, Step 5 is mandatory only for cross-cultural research where different languages are used. 16.9% of measures have reported on translation and back-translation of items or questionnaires.

Before evaluating the new measure on a large sample, it is suggested to conduct a preliminary study on a moderately sized sample on the targeted population (Clark & Watson, 1995; Netemeyer et al., 2003). Eight measures have undergone this step. The sampling and data collection step has a low percentage (16.9%) since sampling and data collection was done for the purpose of the whole study and mostly not for the purpose of testing the measure before conducting the study as a whole.

Finally, the reliability assessment step was accomplished in 90.9% of cases. Although the result is acceptable, it is not yet satisfactory. Accordingly, measures can be reliable without being valid, but cannot be valid without being reliable (Chandler & Lyon, 2001). The study revealed that only 26.0% and 36.4% of measures have been evaluated for dimensionality and construct validity, respectively. But there is consensus that no measure is useful in the long run without evidence of its validity (Nunnally & Bernstein, 1994).

Overall, there are two main shortcomings in regard to the measurement issue in entrepreneurship research that spring out of our study. First, we noted that all too often, specifically in 32,7% of cases, inappropriate scales to measure constructs were used and this refers to the *neither referenced nor developed measures*. This for sure represents a relatively high portion of measures for an emerging but growing field that tries to get legitimacy. Despite that there is also an encouraging portion of measures that are correctly referenced and applied and this includes *referenced* and *adapted measures* in the amount of 29,1% of all measures reported. Secondly, the scale development review seems even more challenging since the majority of steps have not been accomplished to a high percentage. The most reassuring result is that researchers are aware that reliability of a measure is really important. Despite that, much work should be done in future research to bring also other important steps in scale development to a satisfactory accomplishment; beginning from specifying what is the measure about, generating a large pool of items and evaluating the measure for its content validity, that is considering and accomplishing the first phase () on which all the subsequent stages in scale development lie.

3. CONCLUSION

With this study we made a further step towards raising the awareness that the methodological rigor in establishing new measures and using previously validated measures is a

crucial predisposition for the development of any research field, and in particular also of the young and growing field of entrepreneurship. By reviewing the type of measures used and the procedures applied when developing new measures in papers published from January 2009 to December 2010 in *Journal of Business Venturing* and *Entrepreneurship Theory & Practice*, we acknowledged two main problematic issues. Among the 1171 measures reported in 98 empirical papers from our study (1) one third (32.7%) of the measures reviewed were neither developed nor previously used and cannot be counted as appropriate since they were *neither referenced nor developed measures* and (2) when developing new measures several steps were omitted and again such *new measures* are problematic. The main argument is that no matter how profound the theoretical formulations, how sophisticated the design, and how elegant the analytic techniques are, they cannot compensate for poor measures (Pedhazur & Pedhazur Schmelkin, 1991). Similarly, Judge et al. (2007) opinion that regardless of the quality of an idea, the ability to draw inferences about a phenomenon is constrained by the quality of the methods used to gather data about it.

Thus, inappropriate measures and poor measurement development threaten the understanding of entrepreneurship phenomena and paradigm development. Accordingly, several scholars agree that appropriate measurement with valid and reliable measures is the foundation of scientific research and progress in science (e.g. Colquitt & Zapata-Phelan, 2007; Crook et al., 2009; DeVellis, 2003; Judge et al., 2007; Kuskova et al., 2011; Netemeyer et al., 2003; C. R. Reynolds, 2010) while acknowledging the importance of theoretical development (Colquitt & Zapata-Phelan, 2007; Dean et al., 2007). In addition to that, research findings can be called into question if the underlying constructs were not measured properly implying that construct measurement is the foundation of quality empirical research (Kerlinger & Lee, 2000). Crook et al. (2010) add that to the extent that a study is not properly designed and when constructs are not properly measured, not only is confidence in the findings of that study limited but also the field's ability to build on it in future studies. For this reason there is growing accordance that empirical studies that are firmly grounded in theory and use valid measures of their focal constructs are more critical to the advancement of science than studies that do not possess these attributes (Judge et al., 2007; Kuskova et al., 2011)

The importance of applying rigorous methodology in general, and specifically in regards to measure use or measure development, is strengthened by studies that analyzed the influence of measurement rigor on citation count (Bergh, Perry, & Hanke, 2006; Colquitt & Zapata-Phelan, 2007; Judge et al., 2007; Kuskova et al., 2011). For example, in their study, Colquitt and Zapata-Phelan (2007) found that articles which rated (on their rating instrument) as moderate to high on theory building and theory testing enjoyed the highest level of citations. Since the number of citation a paper will get is also predicted by the quality of the journal in which the paper is published (meaning that a high journal average citation rate has a positive influence on the number of citations a paper will get (Judge et al., 2007)), researchers should put great effort into sound measurement development to publish in top-tier journals and get cited more.

To the extent that entrepreneurship is a young and developing research field (Dean et al., 2007; Shane & Venkataraman, 2000) with methodological rigorousness to be improved (Crook et al., 2010; Low, 2001) it is not superfluous to emphasize also with the results of our study that sustained efforts should be undertaken to reach an adequate level of methodological rigorous. All the more is this true for the young discipline of entrepreneurship since it has been argued that relatively new disciplines tend to have less clearly defined theoretical and methodological paradigms than more mature disciplines (e.g. Kuskova et al., 2011; Pfeffer, 1993). And both, theory building and theory testing represent key components of theoretical contribution that can coexist within a given stream of research (Colquitt & Zapata-Phelan, 2007).

Based on the findings of our study we can conclude, similarly to Crook et al (Crook et al., 2010), that we are not there yet. Therefore, it is important that researchers use validated scales or to develop new scales according to an established procedure. To get an overview of the crucial steps in scale development in entrepreneurship research we propose a ten-step methodology in the Appendix of this paper. Although there is no definitive checklist and that procedures vary among authors, the ten steps summarize all crucial steps for developing new measures in entrepreneurship research. While the ten steps are in part tailored for entrepreneurship research, they are consistent with the main-stream recommendations for developing new measures (e.g. Bagozzi & Edwards, 1998; Carmines & Zeller, 1979; Churchill, 1979; DeVellis, 2003; Hinkin, 1998; Netemeyer et al., 2003; Nunnally & Bernstein, 1994; Pedhazur & Pedhazur Schmelkin, 1991) and for the interested readers we suggest consulting also those readings.

In short, the ten-step procedure that is grouped into three phases is the following. The first phase deals with the theoretical importance and existence of the construct and is comprised by the first three steps. The first step in developing a new construct is to specify the content domain of the construct and to propose its definition. In the second step an item pool that samples the domain of the new construct is generated. The third step regards the evaluation of content validity by assessing the relevance of items by the relevant audience, e.g. entrepreneurs, entrepreneurship experts, and scholars. Here, the second phase on the representativeness and appropriateness of data collection begins and is comprised by steps 4 to 7. In the fourth step the questionnaire is developed and evaluated. The fifth step is mandatory for entrepreneurial cross-cultural studies in different languages since the translation and back-translation of the questionnaire is needed. In the sixth step a pilot study is conducted on a relevant audience, presumably entrepreneurs. Then, in the seventh step random sampling on a dataset of entrepreneurs and data collection occur with established tailored design methods for improving response rates. The final three steps represent the third phase on the statistical analysis and statistical evidence of the construct and involve assessment of dimensionality, reliability, and construct validity of the new entrepreneurship-related measure.

We hope these guidelines will contribute to a better and more effective approach for introducing new constructs in entrepreneurship domain and to the advancement of the entrepreneurship field. In so doing, quality measurement design will reach the rigor of

empirical research as it is in other fields that several scholars of entrepreneurship called for (e.g. Chandler & Lyon, 2001; Crook et al., 2009; Crook et al., 2010; Dean et al., 2007; Low, 2001; Low & MacMillan, 1988).

3.1. Limitations and future research avenues

There are two main limitations of our study. The first concerns the limited number of journals reviewed. It might be that reviewing a larger variety of journals may lead to slightly different results. Despite this, we reviewed the two highest-quality entrepreneurship journals and it is quite likely that reviewing lesser quality journals (Crook et al., 2010) would reveal major measurement problems. Accordingly, it has been shown that higher quality journals tend to publish articles that possess more rigorous methods (Kuskova et al., 2011). Secondly, it might be the case that researchers used referenced measures (and not *neither referenced nor developed measure*) or that steps in scale development of *new measures* have been accomplished but the papers did not report on that. Unfortunately, we were not able to control for these shortcomings. This leads us to stressing the importance of consistent reporting of scale development steps, measures that were applied in studies and on other measurement-related steps. Therefore, a further implication attained from our study is that researchers in the field of entrepreneurship need to provide more precise descriptions of measures used or descriptions on the operationalization of their new measures. All in all we call for a valid use of established measures and rigorous development of new measures.

REFERENCES

Bagozzi, R. P., & Edwards, J. (1998). A general approach for representing constructs in organizational research. *Organizational Research Methods,* 1 (1), 45-87.

Bagozzi, R. P., Yi, Y., & Phillips, L. W. (1991). Assessing construct validity in organizational research. *Administrative Science Quarterly,* 36 (3), 421.

Bergh, D. D., Perry, J., & Hanke, R. (2006). Some predictors of SMJ article impact. *Strategic Management Journal,* 27 (1), 81-100.

Brislin, R. W. (1970). Back-translation for cross-cultural research. *Journal of Cross-Cultural Psychology,* 1 (3), 185-216.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin,* 56 (2), 81-105.

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment* (Vol. 17). Newbury Park, CA: Sage Publications, Inc.

Chandler, G. N., & Lyon, D. W. (2001). Issues of research design and construct measurement in entrepreneurship research: The past decade. *Entrepreneurship: Theory & Practice,* 25 (4), 101-113.

Churchill, G. A. J. (1979). A paradigm for developing better measures of marketing constructs. *Journal of Marketing Research,* 16 (2), 64-73.

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment,* 7 (3), 309-319.

Collins, D. (2003). Pretesting survey instruments: An overview of cognitive methods. *Quality of Life Research,* 12 (3), 229.

Colquitt, J. A., & Zapata-Phelan, C. P. (2007). Trends in theory building and theory testing: A five-decade study of the Academy of Management Journal. *Academy of Management Journal,* 50 (6), 1281-1303.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika,* 16 (3), 297-334.

Crook, T. R. et al. (2009). A review of current construct measurement in entrepreneurship. *International Entrepreneurship and Management Journal,* 6 (4), 387-398.

Crook, T. R. et al. (2010). Are we there yet? An assessment of research design and construct measurement practices in entrepreneurship research. *Organizational Research Methods,* 13 (1), 192-206.

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Joumal of Consulting Psychology,* 24, 349-354.

Dean, M. A., Shook, C. L., & Payne, G. T. (2007). The past, present, and future of entrepreneurship research: Data analytic trends and training. *Entrepreneurship: Theory & Practice,* 31 (4), 601-618.

DeVellis, R. F. (2003). *Scale development: theory and applications* (2nd ed. Vol. 26). Thousand Oaks, CA: Sage Publications.

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2009). *Internet, mail, and mixed-mode surveys: the tailored design method* (3rd ed.). Hoboken, N.J.: John Wiley & Sons.

Flynn, B. B., Schroeder, R. G., & Sakakibara, S. (1994). A framework for quality management research and an associated measurement instrument. *Journal of Operations Management,* 11 (4), 339-575.

Hair, J. F. et al. (2010). *Multivariate data analysis* (7th ed.). Upper Saddle River, NJ: Prentice-Hall.

Hardesty, D. M., & Bearden, W. O. (2004). The use of expert judges in scale development: Implications for improving face validity of measures of unobservable constructs. *Journal of Business Research,* 57 (2), 98-107.

Haynes, S. N., Richard, D. C. S., & Kubany, E. S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment,* 7 (3), 238-247.

Hinkin, T. R. (1998). A brief tutorial on the development of measures for use in survey questionnaires. *Organizational Research Methods,* 1 (1), 104-121.

Hui, C. H., & Triandis, H. C. (1985). Measurement in cross-cultural psychology: A review and comparison of strategies. *Journal of Cross-Cultural Psychology,* 16 (2), 131-152.

Jackson, D. N. (1971). The dynamics of structured personality tests. *Psychological Review,* 78 (3), 229-248.

Judge, T. A. et al. (2007). What causes a management article to be cited—Article, author, or journal? . *Academy of Management Journal,* 50 (3), 491-506.

Kerlinger, F. N., & Lee, H. B. (2000). *Foundations of behavioral research* (4th ed.). Fort Worth, TX: Harcourt College Publishers.

Kuskova, V. V., Podsakoff, N. P., & Podsakoff, P. M. (2011). Effects of theoretical contribution, methodological rigor, and journal quality, on the impact of scale development articles in the field of entrepreneurship. *Strategic Entrepreneurship Journal,* 5 (1), 10-36.

Leavitt, C., & Walton, J. (1975). Development of a scale for innovativeness. *Advances in Consumer Research,* 2, 545-554.

Loevinger, J. (1957). Objecive tests as instruments of psychological theory. *Psychological Reports,* 3, 635-694.

Low, M. B. (2001). The adolescence of entrepreneurship research: Specification of purpose. *Entrepreneurship: Theory & Practice,* 25 (4), 17-25.

Low, M. B., & MacMillan, I. C. (1988). Entrepreneurship: Past research and future challenges. *Journal of Management,* 14 (2), 139-161.

Netemeyer, R. G., Bearden, W. O., & Sharma, S. (2003). *Scaling procedures: Issues and applications*. Thousand Oaks, CA: Sage Publications.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw Hill, Inc.

Pedhazur, E. J., & Pedhazur Schmelkin, L. (1991). *Measurement, design, and analysis: An integrated approach*. Hillsdale, New Jersey: Lawrence Erlbaum Associates Inc., Publishers.

Pfeffer, J. (1993). Barriers to the advance of organizational science: Paradigm development as a dependent variable. *The Academy of Management Review,* 18 (4), 599-620.

Presser, S., & Blair, J. (1994). Survey pretesting: Do different methods produce different results? *Sociological Methodology,* 24, 73-104.

Presser, S. et al. (2004). Methods for testing and evaluating survey questions. *The Public Opinion Quarterly,* 68 (1), 109-130.

Reynolds, C. R. (2010). Measurement and assessment: An editorial view. *Psychological Assessment,* 22 (1), 1-4.

Reynolds, W. M. (1982). Development of reliable and valid short forms of the Marlowe-Crowne social desirability scale. *Journal of Clinical Psychology,* 38 (1), 119-125.

Schwab, D. P. (1980). Construct validity in organization behavior. In Staw, B. M. & Cummings, L. L. (Eds.), *Research in Organizational Behavior* (Vol. 2, pp. 3-43). Greenwich, CT: JAI Press Inc.

Shane, S., & Venkataraman, S. (2000). The promise of entrepreneurship as a field of research. *Academy of Management Review,* 25 (1), 217-226.

Strahan, R., & Gerbasi, K. C. (1972). Short, homogeneous versions of the Marlow-Crowne social desirability scale. *Journal of Clinical Psychology,* 28 (2), 191-193.

Zaichkowsky, J. L. (1985). Measuring the involvement construct. *Journal of Consumer Research,* 12 (3), 341-352.

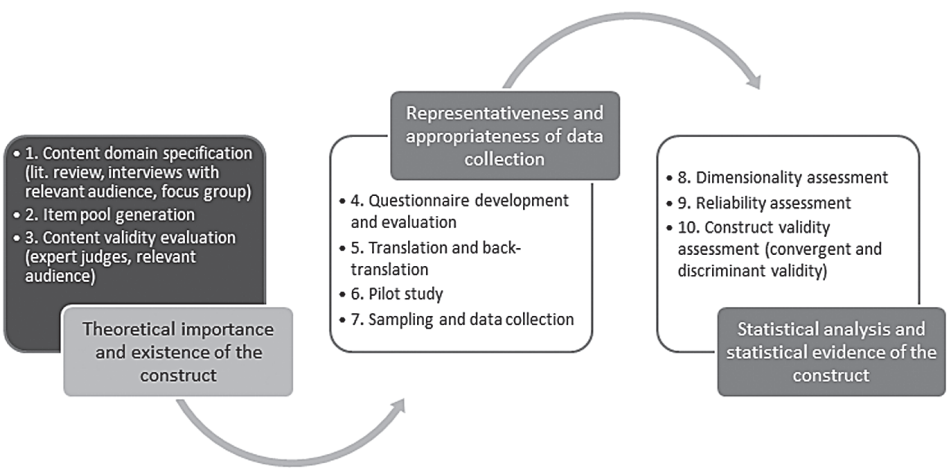
APPENDIX

Guidelines for scale development in entrepreneurship research: A ten-step procedure

To strengthen the measurement part of entrepreneurship-related papers we propose a ten-step procedure for scale development in entrepreneurship research. While our guidelines are tailored for entrepreneurship research, they are consistent with the mainstream recommendations for developing new measures (e.g. Bagozzi & Edwards, 1998; Carmines & Zeller, 1979; Churchill, 1979; DeVellis, 2003; Hinkin, 1998; Netemeyer et al., 2003; Nunnally & Bernstein, 1994; Pedhazur & Pedhazur Schmelkin, 1991). We grouped the ten steps of scale development into three phases: (1) theoretical importance and existence of the construct, (2) representativeness and appropriateness of data collection, and (3) statistical analysis and statistical evidence of the construct. The three phases with the ten steps are discussed in what follows.

Phase 1: Theoretical importance and existence of the construct
The first phase regards the theoretical importance and existence of the construct and is comprised by the following three steps: (1) content domain specification, (2) item pool generation, and (3) content validity evaluation. Figure 3 represents the scale development procedure with the first phase marked in different color.

Figure 3: *Phase 1 – Theoretical importance and existence of the construct*



Step 1: Content domain specification

The first step in constructing a new measure is to clearly define what the researcher intends to measure (DeVellis, 2003). Accordingly, in the entrepreneurial context it should be clearly stated what is the new entrepreneurial construct about, what is to be included in the construct and not less importantly also what is to be excluded from the construct.

Netemeyer et al. (2003) explain that it is crucial to assess the boundaries of the construct domain and not to include extraneous factors or domains of other constructs. Additionally, the domain of the construct should not be set to narrowly and fail to include important facets of the construct. The researcher should also explain how the new construct is different from other related construct that already exist. At this stage an initial definition of the new construct should be proposed (Nunnally & Bernstein, 1994). This working definition might be modified during the scale development process. Besides the definition, it is important to specify dimensions of the new construct (Haynes et al., 1995).

Consistent with scale development experts (e.g. DeVellis, 2003; Netemeyer et al., 2003), we propose that this step is to be achieved by an in-depth interdisciplinary literature review. The review of the relevant literature serves as the basis for grounding the theory of the new construct. The review should include previous attempts to conceptualize and assess both the same construct and closely related constructs (Clark & Watson, 1995). Additionally, we suggest that in entrepreneurship research the existence of the new construct should be explored also in practice. At the very beginning of scale development, interviews with the relevant audience from entrepreneurship should be conducted. We suggest conducting semi-structured interviews on the topic of the target construct with entrepreneurs, entrepreneurship professors, business angels, venture capitalists, representatives of technology parks, business incubators, chambers of trade, and financial institutions that deal with entrepreneurs and the like. These interviews should evidence the importance and the need for the new measure. Interviews should be taped, transcribed, and analyzed in order to get feedback on eventual inclusion of some aspects of construct domain that might have not been included. Similarly, suggestions on exclusion of some other parts of the domain might be raised. We also propose conducting a focus group with 5-7 entrepreneurial experts, practitioners or other relevant profiles that would brainstorm, confront and upgrade ideas and views on the target construct. Like the interviews, the focus group should be taped, transcribed, and analyzed. Interviews and focus groups can be effectuated in several rounds if needed. After each round, relevant literature should be reviewed again for parts that might have been missing. In each round, an improved definition is proposed and interviews become more structured. A researcher should also analyze interviews, web pages, statements and lectures of entrepreneurial scholars, entrepreneurs, venture capitalists, business angels, etc. and consult popularly-scientific media on the target topic to gather additional insights.

At the final stage of content domain specification, the researcher should be able to clearly state why the new measure is needed given that several entrepreneurial constructs exist. The explanation should contain evidence that the older measures, if they exist, do not satisfy the research needs or that inadequate measures have been used. It may also be the case that no previous measure has been proposed for the targeted entrepreneurial construct although evidence from the literature and the field suggests its importance.

A well-expressed theoretical base that clearly specifies the content domain of the construct is crucial for all subsequent steps (Netemeyer et al., 2003). Based on the specified domain of the construct the generation of an item pool takes place.

Step 2: Item pool generation

After specifying the domain and defining the purpose of the construct a researcher generates a large pool of items that are candidates for eventual inclusion in the scale (DeVellis, 2003) and that captured the domain (Churchill, 1979) of the new entrepreneurial construct. An initial list of items can be generated applying techniques that are usually productive in exploratory research and include a literature review, examination of existing scales related to the new construct (Churchill, 1979) and expert judges (DeVellis, 2003; Hardesty & Bearden, 2004). Additional items can be created based on interviews and focus groups conducted in the first step. A researcher can get ideas for item development also from the analysis of interviews, web pages, statements and lectures of entrepreneurial expert and practitioners effectuated in the first step as well as from the popularly-scientific press.

Writing items is an art (Nunnally & Bernstein, 1994). However, good items can be written following some basic rules (e.g. Clark & Watson, 1995; DeVellis, 2003; Netemeyer et al., 2003; Nunnally & Bernstein, 1994). Items should be clear, simple and short. A researcher should pay attention also to the reading level ease of the target population, e.g. entrepreneurs and entrepreneurial students, should avoid multiple negatives, double barreled sentences, and use of jargon or trendy expressions. Items should be written in a way to ensure variability in responding. The initial item pool should include a large number of items since at this stage oveinclusiveness is preferred to underinclusiveness. DeVellis (2003) argues that it would not be unusual to begin with a pool of items that is three to four times as large as the final scale.

Items serve as evidence of content validity. The appropriateness of items and their relatedness to the specified construct is assessed in the next step in which the evaluation of content validity begins.

Step 3: Content validity evaluation

Content validity refers to the degree to which elements of a measurement instrument are relevant to and representative of the targeted construct for a particular assessment purpose (Haynes et al., 1995). Elements of a measurement instrument include items, response formats and instructions. These elements represent key parts of a questionnaire and questionnaires are a widely used data collecting format in entrepreneurship research (Chandler & Lyon, 2001; Crook et al., 2010). Based on its importance, the development of the questionnaire is discussed in the next step. Here we expand on content validity evaluation.

The representativeness criterion has two meanings. First, it refers to the degree to which the domain has been sampled, i.e. if the elements of the construct are proportional to the facets of the target construct (Haynes et al., 1995; Loevinger, 1957; Netemeyer et al., 2003). Accordingly, Nunnally & Bernstein (1994) argue that item pools should be

regarded as samples of content and they should be evaluated in terms of how well they sample the implied domain and how relevant are they for the target construct (DeVellis, 2003). Thus, all items should be reviewed by judges (Hardesty & Bearden, 2004), i.e. knowledgeable people in the content area (DeVellis, 2003) to check for their content validity. Reviewers of items should be chosen broadly (Nunnally & Bernstein, 1994). In an entrepreneurship context we should include end users (scholars that would administer items of the new construct), subject-matter experts (e.g. entrepreneurship professors, business angels, venture capitalists, representatives of technology parks, business incubators, chambers of trade, and financial institutions), psychometricians, and representatives of those who will be taking the test (e.g. entrepreneurs, entrepreneurship students, nascent entrepreneurs). At this stage suggestions for modification, addition, and exclusion of items are often made. The evaluation of items should be in a formal note. Zaichkowsky (1985) proposed an evaluation method, whereby each item is rated as "clearly representative", "somewhat representative" or "not representative" of the new construct.

The second meaning of the representativeness criterion refers to the degree to which elements are representative of the construct (Netemeyer et al., 2003). Thus, it is recommended to take notes of all the comments that judges express in regard to representativeness, clarity and wording of items, clarity of instruction, response formats, and sequence, length, and appearance of the assessment instrument. Based on these comments and the evaluation of the representativeness of items, the measurement instrument should be revised. Despite experts` opinions on retention, modification or exclusion of items, DeVellis (2003) warns that the final decision to accept or reject the advice of experts is the responsibility of the scale developer.

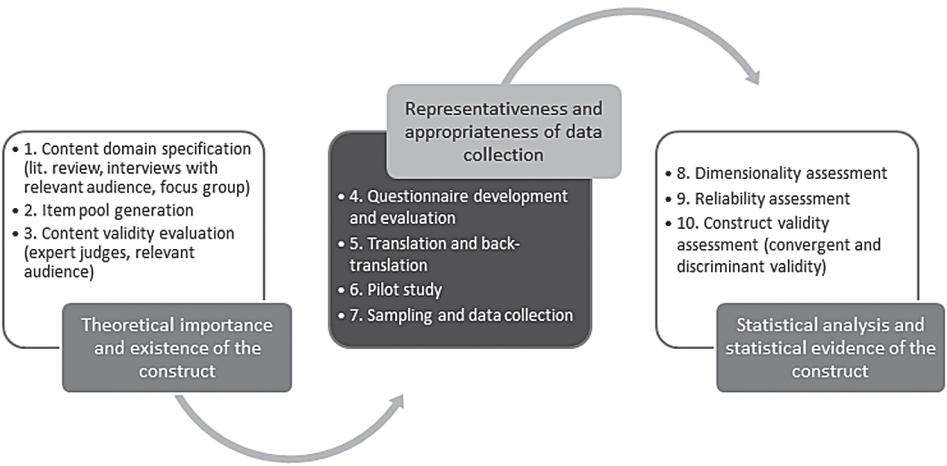
In order to increase the legitimacy of the new construct, information regarding the representativeness, relevance, and evaluation of the measurement instrument should be reported in papers. Besides for the publication reason, tracking all the stages (generation, modification, addition, and elimination of items) is useful also for researchers themselves. It should be noted that the evaluation stage should be repeated as long as items and other elements of the assessment instrument are seen as the most appropriate. This may involve several periods of item writing, followed in each case by their analysis (Clark & Watson, 1995).

These first three steps are crucial in scale development since they influence construct validity. A researcher might not be able to assess construct validity if a construct is not precisely or appropriately defined or if measurement items are not sufficiently representative or relevant of the new construct. C. R. Reynolds (2010) warns that it is not appropriate to simply write a set of items and begin research on a construct since this omits key first steps, namely conducting the necessary research on the measurement instrument itself, modifying it appropriately, trying out the items again – and again if necessary – and then applying the instrument to the research on the construct in question and making it available to other researchers.

Phase 2: Representativeness and appropriateness of data collection

The second phase regards the representativeness and appropriateness of data collection and is comprised by the four steps that follow the previous three steps: (4) questionnaire development and evaluation, (5) translation and back-translation of the questionnaire, (6) pilot study performance, and (7) sampling and data collection. Figure 4 represents the scale development procedure with the second phase marked in different color.

Figure 4: *Phase 2 – Representativeness and appropriateness of data collection*



Step 4: Questionnaire development and evaluation

A common method for data collection in entrepreneurship research is via questionnaires (Chandler & Lyon, 2001; Crook et al., 2010). Besides the commonly used paper-and-pencil version, the usage of web-based surveys is increasing. A properly designed questionnaire facilitates gathering relevant data for the study. Dillman et al. (2009) proposed the tailored design method for internet, mail, and mixed-mode surveys. Following rules for establishing trust, increasing the benefits of participation, and decreasing costs of participation a researcher gets more reliable data and improves response rates. Examples of these rules are: ensure confidentiality, provide information about the survey, ask for help, say thank you, make it convenient to respond, and send remainders (Dillman et al., 2009).

Pre-testing is generally agreed to be an indispensable stage in survey development (Collins, 2003; Dillman et al., 2009; Presser & Blair, 1994; Presser et al., 2004), thus the survey should be pre-tested with people who have specialized knowledge on some aspects of questionnaire quality (Dillman et al., 2009). In the entrepreneurship research setting, we suggest also to administer the pilot version to the target population. This can be carried out as a "think aloud" protocol where the researcher takes notes of comments raised by the evaluator.

When developing the measurement instrument that better suits the targeted construct, e.g. a questionnaire, we suggest including scales of validated related constructs for an initial comparison among constructs already in the pilot testing stage, that is presented in Step 6. Besides validated scales of related constructs, we suggest also the inclusion of a social desirability scale. This serves to detect potential distorted item responses due to the inclination of individuals to represent themselves in a way that society regards as positive (DeVellis, 2003). The most widely used social desirability scale is the Marlowe-Crowne social desirability scale (Crowne & Marlowe, 1960). Since it contains 33 items, several shorter versions have been developed, e.g. a 10-item measure by Strahan and Gerbasi (1972) and a 13-item measure by W. M. Reynolds (1982). With the inclusion of a social desirability scale the Differential Reliability Index can be calculated in order to maximize the content saturation of the item in relation to its saturation with a desirability scale (Jackson, 1971).

Step 5: Translation and back-translation of the questionnaire

The cross-cultural equivalence is a prerequisite for comparison across cultural and ethnic boundaries (Hui & Triandis, 1985). Brislin (1970) suggested several methods aimed at putting the same test in different languages while preserving the same ideas across the linguistic boundaries. One of these methods is back-translation by which a measurement instrument should be translated and back-translated in the original language. The equivalency of the original version and back-translated version should be evaluated.

Step 6: Pilot study performance

At this stage the first data collection and evaluation begins. Before engaging in the large data collection for the assessment of psychometric properties of the new construct, it is important to conduct a preliminary study on a moderately sized sample on the targeted population, e.g. on 30-50 entrepreneurs. Although students are a convenient sample for preliminary research (Netemeyer et al., 2003), a researcher cannot always draw reliable conclusions from their results. The reason is that the construct may have different properties in different samples (Clark & Watson, 1995). C. R. Reynolds (2010) also argues that collegiate samples, especially from a single university or campus, are rarely representative of anything beyond students on that campus. Thus, we suggest conducting the pilot study on the target population, i.e. entrepreneurs, entrepreneurial ventures and others. Entrepreneurship students would be a relevant sample if a research tries to evaluate entrepreneurial intentions or other pre-venturing factors.

This mini-study has two purposes. On the one hand, with a pilot study on the target population a researcher tests the proposed questionnaire in order to identify potential problems with the questionnaire (Dillman et al., 2009). On the other hand, a researcher obtains a pre-result of the reliability of the new measure (Netemeyer et al., 2003) and

examines the boundaries of the target construct and correlations of the target construct with existing measures (Clark & Watson, 1995). In doing so, a researcher is able to detect flaws or problems with the measure and hopefully demonstrate that the new construct is distinct from other related constructs. Clark and Watson (1995) note that all too often researchers discover late in the process that their new measure correlates too highly with an existing measure.

Step 7: Sampling and data collection

With a sampling process we aim at obtaining a representative portion of some whole to afford valid inferences and generalizations to it (Pedhazur & Pedhazur Schmelkin, 1991). Thus, it is important to carefully undergo the sampling process. At this stage the researcher aims at collecting data for the evaluation of the measure`s factor structure and for the subsequent analysis of various types of validity (Hinkin, 1998). Random sampling is the most common type of sampling used in self-administered surveys (Dillman et al., 2009). In random sampling every member of the sample frame has an equal chance of being selected.

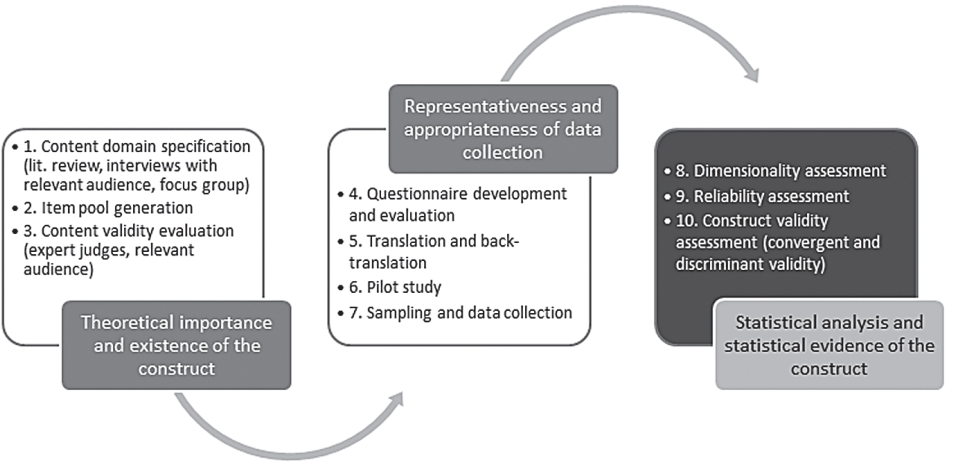
At this stage it is crucial to administer the questionnaire to the actual population of interest (Hinkin, 1998). Thus, the researcher should choose respondents from a database of entrepreneurs or entrepreneurial ventures. Hinkin (1998) also proposes that when developing a scale, it is necessary to use several independent samples. Besides sampling in one country, cross-cultural samples should be taken into account.

While an adequate representation of the target population in the sample is crucial for inferences to the whole target population, also the size of the sample should be considered. There is no established rule about the size of the sample. Still, two main recommendations should be followed. First, the sample of subjects should be large (DeVellis, 2003) and second, as the number of items increases, the number of respondents should increase too (Hinkin, 1998). Some recommendations for item-to-response ratios suggest a 1:10 ratio (Schwab, 1980), by which 100 respondents are needed for a 10-item measure. Similarly, DeVellis (2003) proposes that if only a single scale is to be extracted from a pool of about 20 items, fewer than 300 subjects might suffice. Hinkin (1998) suggests that for factor analysis a sample of at least 200 respondents is required.

Phase 3: Statistical analysis and statistical evidence of the construct

In the final phase statistical analysis and statistical evidence of the construct is performed. This is achieved by the last three steps: (8) dimensionality assessment, (9) reliability assessment, and (10) construct validity assessment. Figure 5 represents the scale development procedure with the third phase marked in different color.

Figure 5: *Phase 3 – Statistical analysis and statistical evidence of the construct*



Step 8: Dimensionality assessment

Dimensionality of a measure is concerned with the homogeneity of items and is defined as the number of common factors needed to account for the correlation among items (Netemeyer et al., 2003). Homogeneity indicates whether the scale items assess a single underlying factor or construct (Clark & Watson, 1995). A measure that is considered unidimensional has a single facet or dimension, which means that its items underlie a single factor. Accordingly, a multidimensional measure has multiple facets or dimensions and its items tap more than one factor (Netemeyer et al., 2003).

Dimensionality can be assessed with either exploratory factor analysis (EFA) or confirmatory factor analysis (CFA) or both. EFA is commonly conducted in the initial stage of scale development. When conducting EFA it is assumed that a researcher has a limited idea of the new measure`s dimensionality (Netemeyer et al., 2003). In CFA the number of factors, the factor structure and the relationship among factors are set a priori. CFA evaluates whether the hypothesized factor model fits the data or not (Netemeyer et al., 2003). Unidimensionality of a scale can be assessed by average interitem correlations taking in consideration also the range and distributions of correlations. Interitem correlations should fall in the range of 0.15 and 0.50 and should cluster narrowly around the mean value (Clark & Watson, 1995).

Step 9: Reliability assessment

After the dimensionality of the new measure has been set, the researcher proceeds with assessing measure`s reliability. Scale reliability is the proportion of variance attributable to the true score of the latent variable (DeVellis, 2003). Thus, it deals with that portion of

measurement that is due to permanent effects that persist from sample to sample (Netemeyer et al., 2003). Scale reliability can be assessed with several methods, e.g. temporal stability, split-half reliability and internal consistency. The temporal stability or test-retest reliability refers to correlations between the same person`s score to the same test at two different points in time (DeVellis, 2003). In split-half reliability the researcher divides the set of items into two subsets and correlates the subsets to assess the reliability (DeVellis, 2003). But the single most widely used method for item selection in scale development is some form of internal consistency analysis (Clark & Watson, 1995). Internal consistency refers to the homogeneity of items within a scale (DeVellis, 2003). Thus, a scale is internally consistent if items of a scale are highly intercorrelated. High inter-item correlations suggest that items are measuring the same construct (DeVellis, 2003). The most commonly employed internal consistency measure is the Cronbach`s (1951) coefficient alpha. Cronbach alpha represents the mean of all split-half coefficients resulting from different splittings of a test (Cronbach, 1951). Different suggestion have been proposed for the acceptable level of coefficient alpha, but an alpha above 0.8 constitutes a reliable measure (Carmines & Zeller, 1979; Clark & Watson, 1995). Besides this, Hair et al. (2010) suggest that the generally agreed upon lower limit for Cronbach's alpha is 0.70, although it may decrease to 0.60 in exploratory research. Moreover, internal consistency can be evaluated also with item-to-total correlations and inter-item correlations. The item-to-total correlation is the correlation of the item to the summated scale score and items with correlations below 0.50 with the scale score should be eliminated (Hair et al., 2010). The inter-item correlation refers to correlations among items and items with correlations below 0.30 with other items should be excluded (Flynn, Schroeder, & Sakakibara, 1994; Hair et al., 2010). When evaluating items, Clark and Watson (1995) propose that items that have highly skewed and unbalanced distributions should be removed, while it is reasonably to retain items that show a broad range of distributions.

If a social desirability scale is included in the study of the new construct, the Differential Reliability Index (Jackson, 1971) can be evaluated. This serves to identify answers to items that are biased by social desirability. It may be that some respondents try to present themselves in a more social desirable way. For each item, the correlation between the scale item and the total desirability score is subtracted from the correlation between that scale item and the total scale score. The square root of the remainder is the proportion of true content variance for any given item (Leavitt & Walton, 1975).

Step 10: Construct validity assessment

In the final step of scale development, validity of the measure is to be assessed. Validity has been given several meanings and consensus has not been reached for the categorization of all the types. Construct validity seems the most important one and for some scholars construct validity subsumes all categories of validity (Haynes et al., 1995; Netemeyer et al., 2003). Construct validity refers to the extent to which any measuring instrument measures what it is intended to measure (Carmines & Zeller, 1979; Netemeyer et al., 2003) in the context in which it is to be applied (Nunnally & Bernstein, 1994). Put

in other words, a researcher measures the extent to which a set of measured items actually reflects the theoretical latent construct those items are designed to measure (Hair et al., 2010). Among other types, construct validity is investigated through convergent and discriminant validity.

Convergent validity is the extent to which responses from alternative measurements of the same construct share variance (Schwab, 1980). That is, a measure has convergent validity if independent measures of the same construct converge, or are highly correlated (Netemeyer et al., 2003). Discriminant validity refers to the degree to which two measures designed to measure similar, but conceptually different constructs are related (Netemeyer et al., 2003).

Convergent and discriminant validity can be assessed with different methods, e.g. by multitrait-multimethod matrix (MMTM matrix) (Campbell & Fiske, 1959; Netemeyer et al., 2003) and by factor analysis. While MMTM is useful in assessing convergent and discriminant validity, it has been criticized by several researchers because of unrealistic assumptions (e.g. Bagozzi, Yi, & Phillips, 1991). Factor analysis, specifically CFA, is strongly recommended at this stage. In CFA a researcher should examine the loadings of items on factors. Items that load on a factor relatively strongly and weakly on other factors are candidates for retention (Clark & Watson, 1995). Factor loadings should be statistically significant and standardized loadings should be above 0.50 (ideally 0.70) (Hair et al., 2010). Additionally, average variance extracted (AVE) can be analyzed. AVE is a summary indicator of convergence. An AVE of 0.50 or higher suggests adequate convergence. On the other hand, discriminant validity refers to the absence of correlation between measures of unrelated constructs (DeVellis, 2003), therefore the presence of cross-loadings would present a discriminant validity problem (Hair et al., 2010). Items that loaded on different factors should be eliminated.