Lost in translation?

An explorative and comparative analysis of the psychometric qualities of the Norwegian Leadership Versatility Index

Sofie Bang Jensen & Ida Kopperstad

Graduate thesis at the Department of Psychology

UNIVERSITY OF OSLO

Spring 2019
Lost in translation?
An explorative and comparative analysis of the psychometric qualities of the Norwegian Leadership Versatility Index

Ida Kopperstad
Sofie Bang Jensen
“I am among those who think that science has great beauty”

- Marie Curie
Abstract

Lost in translation? An explorative and comparative analysis of the psychometric qualities of the Norwegian Leadership Versatility Index
By Sofie Bang Jensen & Ida Kopperstad
Supervisors: Henning Bang & Dag-Erik Eilertsen

Flexible or versatile leadership has been found to be essential to performance in organizations. Versatile leadership is defined as adapting leadership behavior to changes in contextual demands, conceptualized as the mastery of opposing but complementary behaviors. The Leadership Versatility Index (LVI) is a multi-rater feedback tool, where a leader’s performance is rated by several coworkers. The tool is designed to help managers understand their leadership behavior along two domains, categorized as how one leads people in the organization and what organizational issues the leader focuses on. This is measured through the leader’s versatility within four dimensions of leadership behaviors: forceful, enabling, operational and strategic leadership.

The English version of the LVI has been extensively researched and validated over the past three decades, and was translated to Norwegian in 2017. The purpose of the present study is to investigate the psychometric qualities of the Norwegian version of the LVI and how it compares to the American version. The thesis utilizes a quantitative design, analyzing ratings of 44 Norwegian leaders recruited for the present project. Measures previously used to assess the reliability and validity of the American LVI are utilized to allow for comparison.

Overall, the results indicate a satisfactory level of reliability and validity for the forceful and enabling dimensions. The strategic-operational dimension pair displayed considerable weaknesses in reliability and internal structure. The underlying theoretical model was not confirmed by factor analysis, but the results could be a product of methodological limitations. In the sample, versatility explained half of the variation in what separates the most effective from the least effective managers. Furthermore, the Norwegian sample differed from the global sample in which behavior dimensions that were most over- and underdone. Except for these differences, the results were to a large part similar to those of the American LVI.
Acknowledgements

This thesis would not have been realized without the amazing guidance we have received from our associate professor, Henning Bang. Thank you for being our supervisor/Yoda and always cheering for us. We also have to thank our statistics guru, associate professor Dag-Erik Eilertsen. Without your thorough explanations of classical test theory and factor analysis, the wondrous world of statistics would still have been out of our reach.

A special thanks goes to Robert Kaiser, whose patience with our many questions regarding the LVI has been both impressive and heart-warming.

We would also like to thank our families and friends for the unending support, love and care they have given us.

Oslo, April 2019

Sofie Bang Jensen & Ida Kopperstad
# Table of Contents

1  Clarifications and Definitions .......................................................... 3
   1.1  Leadership .................................................................................. 3
   1.1.1  Flexible leadership .................................................................. 3
   1.1.2  The how and what of leadership ............................................. 4
   1.2  The LVI structure ....................................................................... 5
   1.2.1  Strengths overused and lopsided leadership .......................... 7
   1.3  Multi-rater assessments ............................................................. 8
   1.4  The LVI scale ........................................................................... 9
   1.4.1  The versatility measures ......................................................... 10
   1.4.2  Effectiveness indicators ......................................................... 10
   1.5  Differences in American and Norwegian professional culture .......... 11

2  Rationale and Research Questions ...................................................... 13
   2.1  Reliability .................................................................................. 13
   2.1.1  Scale reliability ..................................................................... 13
   2.1.2  Inter-rater reliability .............................................................. 13
   2.1.3  Inter-rater agreement .............................................................. 13
   2.2  Validity ....................................................................................... 14
   2.2.1  Content Validity .................................................................... 14
   2.2.2  Relationship with effectiveness indicators ............................ 16
   2.2.3  Differences in frequency of leadership behaviors .................. 17

3  Method ............................................................................................ 18
   3.1  Administration .......................................................................... 18
   3.2  Sample ....................................................................................... 18
   3.3  Ethical and legal considerations .................................................. 19
   3.4  The Norwegian translation ......................................................... 19
   3.5  Level of analysis ...................................................................... 19

4  Results ............................................................................................ 20
   4.1  Descriptive statistics ................................................................. 20
   4.2  Reliability .................................................................................. 23
   4.2.1  Scale reliability ..................................................................... 23
   4.2.2  Inter-rater reliability .............................................................. 24
   4.2.3  Inter-rater agreement .............................................................. 26
   4.3  Validity ....................................................................................... 27
   4.3.1  Content validity - Relationship between opposites ................ 27
4.3.2 Content validity - Exploration of the conceptual structure 29

4.4 The Norwegian LVI and effectiveness indicators 35
4.4.1 The relationship between perceived leader effectiveness and the four dimensions 35
4.4.2 The relationship between perceived leader effectiveness and versatility 36
4.4.3 Which form of versatility is the stronger predictor? 37
4.4.4 The relationship between versatility and effectiveness indicators 38

4.5 Differences in the frequency of leadership behaviors 41

5 Discussion 43
5.1 Theoretical Implications 44
5.1.1 The operational leadership dimension 44
5.1.2 Are strategic and operational opposites, or simply complementary? 45
5.1.3 Factor analyses and the underlying theoretical model 45
5.1.4 Effectiveness indicators 47
5.1.5 Is there an underlying trait behind the two forms of versatility? 47
5.1.6 Differences in degree of overdone and underdone behaviors 48
5.1.7 Effects of femininity on correlations between versatility and effectiveness indicators 49

5.2 Limitations and future research 49
5.2.1 Sample 49
5.2.2 Public sector 50
5.2.3 Translation 50
5.2.4 Restriction of range 50
5.2.5 Common method bias 51
5.2.6 Sampling 52
5.2.7 Recommendations for future research 52

5.3 Practical Implications 53

6 Conclusion 54

7 References 55

8 Appendix 65
Appendix 1 65
Appendix 2 66
Appendix 3 67
Appendix 4 69
Appendix 5 70
Appendix 6 71
Lost in translation? An explorative and comparative analysis of the psychometric qualities of the Norwegian Leadership Versatility Index

In a study on the Norwegian work-force, about a third reported that they were “often” exposed to destructive leadership in some form (Aasland, Skogstad, Notelaers, Nielsen, & Einarsen, 2010). Multiple studies have shown that leaders can have a substantial effect on several different areas and aspects of an organization (Aasland et al., 2010; Kaiser & Overfield, 2010; Yukl & Mahsud, 2010; Yukl, 2013). This highlights the need for tools that can help leaders reach their maximum potential, for the betterment of businesses as well as the well-being of the majority of the adult population. The American multi-rater assessment tool Leadership Versatility Index (LVI) claims to be such a tool, a claim supported by almost three decades of research (e.g. Kaiser, McGinnis & Overfield, 2012; Kaplan & Kaiser, 2010). The unique value proposition of the LVI is that it captures the balancing acts and trade-offs that are an essential part of being a leader. This is achieved through a rating scale that measures both when leaders are overdoing and when they are underdoing behaviors, as well as through a dimensional model of opposing, but complementary behaviors.

With the term “opposing but complementary”, it is meant that behaviors which seem opposite - e.g. to take charge versus empower your employees, or to focus on organizational efficiency versus growth - are still complementary behaviors that should be balanced across time and situations. Versatile leaders are expected to meet these tensions and contradictions inherent in the leader role by optimal use of different leadership behaviors, in response to the complex contextual demands in the world of modern businesses (Kaplan & Kaiser, 2013). This should be reflected in an overall leader behavior style perceived as neither "too much" nor "too little" on any of the opposing behaviors, but rather as the “right amount”, reflected in the versatility score in the LVI.

All the research on the LVI so far has been on the English version. Recently, a Norwegian translation was completed, which as of yet has not been evaluated empirically. Although it is not unreasonable to assume that a tool developed in American culture will work for Norwegian populations as well, there can be significant differences in organizational culture that make the assessment tool show different psychometric qualities in a Norwegian setting compared to an American (e.g. Dickson, Hartog & Mitchelson, 2003; Hofstede, 2011). Translation can also have unintended effects on the quality of an assessment tool (Cha, Kim & Erlen, 2007). The purpose of the present study is therefore to examine the psychometric properties of the
Norwegian version of LVI, and to what extent the psychometric properties of the Norwegian LVI resembles the American version.

If the results from the following analyses show promise, it would serve as an indication that the Norwegian LVI could be a valid and reliable tool for measuring important leadership behaviors. The data can also serve as a base for continued research and development of the assessment tool.
1 Clarifications and Definitions

1.1 Leadership

Leadership is a complex phenomenon, which can be defined as “a process whereby an individual influences a group of individuals to achieve a common goal” (Northouse, 2010, p. 5). Throughout the 20th and 21st century, various theories have contributed to the field with different perspectives on leadership: from the first trait-based theories, to theories on skill and behavior, to the more recent situation- and contingency-oriented theories (e.g. Northouse, 2010; Yukl, 2013). Over the years, the focus in leadership research has shifted from the leader alone to exploring the relationship between leader and member (e.g. Leader-member exchange theory, Graen & Uhl-Bien, 1995). Finally, the current trend in leadership theory is focused on inspiration and change (Yukl, 2013). The diversity of leadership as concept illustrates the intricacy of the phenomenon, influenced by multiple aspects of the leader, the team, the organization and even the society and culture in which leadership occurs.

1.1.1 Flexible leadership

The concept of flexible leadership has received a lot of attention in recent years (e.g. Goldsmith & Reiter, 2007; Yukl & Mahsud, 2010). Central to all definitions of the construct is the idea that leaders should change in the way they behave, react and make decisions according to changes in circumstance (Norton, 2010). Kaiser and Overfield define flexible leadership as “adjusting one’s leadership style, method or approach in response to different or changing contextual demands in a way that facilitates group performance” (Kaiser & Overfield, 2010, p.106) and calls it versatile leadership. This implies mastery of a broad spectrum of leadership behaviors, as well as knowing when and how to implement them (Kaplan & Kaiser, 2003). Additionally, Kaiser and Overfield (2010) emphasize a focus on positive outcomes of the leader behavior - specifically within the areas of perceived leader effectiveness, team productivity and team vitality - meant to represent different effectiveness indicators.

Although versatility in leadership behavior correlates with the personality trait of flexibility, previous studies have found the two to be distinct (Lindberg & Kaiser, 2004). In comparing the different methods of measuring flexibility, Kaiser, Lindeberg and Craig (2007) also found that the versatility measures in the LVI captures more of what is considered an effective leader than flexibility as a personality trait (Kaiser et al., 2007).
The idea that behavioral flexibility is a central skill for leaders is supported by current research. The most effective leaders are those who develop and execute a broad repertoire of skills and perspectives (Lombardo & Eichinger, 2006; McCall, Lombardo & Morrison, 1988), and flexible leadership has been found to be essential to adaptability and performance in organizations (Kaiser & Overfield, 2010). Several lines of study have also demonstrated that the relevance of different skills depends on the situation: managerial level, type of organization and external environment will all affect how much a behavior is needed (e.g. Gentry, Harris, Baker and Leslie, 2008; Quinn, 1992). This suggests that an essential part of being a successful leader is the ability to adapt your leadership behavior to changing requirements, in accordance with the theoretical model of the LVI.

1.1.2 The how and what of leadership

Based on extensive research and thousands of assessments of individual executives, Kaiser, Overfield and Kaplan (2010) pose that leadership can be divided into two domains: a relations-oriented focus grounded in psychological theories on interpersonal behaviors, and a business-oriented focus that emphasizes organizational functions of leadership. They label these two behavioral categories the how and what of leadership, respectively (Kaiser et al., 2012).

The categorization is supported by several findings, such as the trend for interpersonal behaviors like taking charge, decisiveness, empowerment and participation to correlate with each other to a greater degree than the organizational-oriented behaviors, e.g. setting direction, introducing change, monitoring performance and allocating resources (Kaiser et al, 2007). Research has also shown that the two behavioral categories are differently related to outcomes, with the interpersonal “how”-behaviors having the greatest impact on employee attitudes and motivation, while the organizational “what” behaviors are most related to overall output (Kaiser & Overfield, 2010).

The distinction between the how and what is similar to distinctions in other leadership theories and models. Kaiser and colleagues (2012) compare it to Dubin’s (1979) distinction of leadership in and of an organization, and to Zaccaro and Horn’s (2003) distinction of interpersonal direct influences versus impersonal indirect influences that concerns establishing direction, policies, plans and goals. Kilburg’s (2012) questions around what an organization is going to do and how it will get it done, is also said to be comparable, where what concerns strategic directions and how concerns steering people towards this. Though there are similarities, the LVI-model is unique as it incorporates opposing behavioral dimensions within the two domains, and Kaiser
and colleagues (2012) propose that the *how* and *what* distinction offers a more inclusive and comprehensive categorization than many other existing theories. They further state that theory, research and practice models that include both *how* and *what* will offer the most holistic approach to leadership development, and describe the LVI-model as such a model (Kaiser et al., 2012).

### 1.2 The LVI structure

In the theoretical LVI model, the *how* and *what* domains are further divided into two leadership dimensions each. The interpersonal *how* is composed of forceful and enabling leadership, while strategic and operational leadership make up the organizational *what* domain in the model. The four dimensions in the LVI represent different aspects of leadership: Forceful leadership is defined as using personal and positional power to push for better performance. Enabling leadership is the ability to engage others and facilitate their contribution. Strategic leadership entails a focus on long-term strategy and positioning the business for future competitiveness. Finally, operational leadership is defined as a focus on short-term goals (Kaiser et al., 2012).

The four dimensions are further divided into three subdimensions, or subscales. These are all comprised of four items each, describing a discrete, observable behavior. The subscales and items are modeled as opposed to one another in the same way as the dimensions (see figure 1). As mentioned earlier, these “opposites” are seen as equally important and complementary behaviors that should be balanced in a leader’s overall behavior patterns over time.
Figure 1. The opposing, complementary dimensions and subdimensions.

For forceful and enabling leadership, the three opposing subscale pairs are; (1) *takes charge - empowers*, which concerns authority and taking control versus giving others autonomy; (2) *declares - listens*, making your position known versus being open to input from others and (3) *pushes - supports*, concerning the orientation of the leader, balancing between setting high standards and giving encouragement. In the strategic-operational domain, the pairs are; (1) *direction - execution*, setting the direction versus focusing on implementation; (2) *growth - efficiency*, seeking possibilities to expand versus conserving and (3) *innovation - order*, the degree of experimentation, represented by supporting change versus using process discipline.

Prior research supports the structure of the four dimension scales and the associated subscales (Kaiser, 2006; Kaiser et al., 2010; Kaiser et al, 2012). Throughout the different studies on the LVI, the forceful-enabling dimension pair has had good results on all estimates of reliability and validity (Kaiser et al., 2010). In this thesis, reliability is defined as to what degree the variance in scores is a result of true variation in the relevant construct and not a result of error variance (Wainer & Thissen, 2001). Validity concerns the degree to which a tool or an operationalization accurately reflects the construct it purports to measure (Throchim, 2006a).

The strategic-operational dimension pair has been less clearly supported in examinations of both reliability and validity: The operational scale has until recently displayed lower alpha values and
the strategic-operational dimension pair is less negatively correlated than expected based on the model (Kaiser et al., 2010). Despite these challenges, strategic-operational versatility was most strongly associated with relevant outcome variables (Kaiser et al., 2012). Research has also demonstrated a robust relationship between versatility and outcome measures such as staff motivation and moral, team productivity and financial results (Kaiser et al., 2010). Finally, examinations of the internal structure have found support for a four-factor solution.

1.2.1 Strengths overused and lopsided leadership

Leadership development has mainly focused on the problems of doing too little of a behavior, thus assuming a linear relationship between the behaviors and effectiveness and neglecting the perils of overdoing leadership behaviors (Yukl, 2013). Overdoing a behavior occurs when a leader pays too much attention to one type of leadership behavior (Kaplan & Kaiser, 2006; Kaiser & Kaplan, 2009). Empirical research shows that leaders who overdo one approach tend to neglect the opposing approach: 97% of leaders who overdo forceful leadership also underdo enabling leadership according to their coworkers, while 94% who overdo operational leadership also underdo strategic leadership (e.g. Kaiser & Kaplan, 2009; Kaiser & Overfield, 2010). Kaiser and Overfield (2011) pose that a possible cause for leaders’ trend to neglect the opposite a commonly used leadership behavior, is the human tendency of grouping behaviors or attributes into “me/my ingroup” or “not-me/my ingroup” and avoiding the “not-me” category in order to behave consistently. This tendency is described in several theories of how we shape our identity, among them symbolic interactionism and social identity theory (Hogg & Vaughan, 2011, 125-131; Marcus & Wulf, 1987; McCall, 2003).

Kaplan and Kaiser (2009) describe the pattern of overdoing strengths at the expense of an opposing but complementary perspective, competency or skill as lopsided leadership. Their research shows that lopsided leaders are less effective, have less productive and cohesive teams, and as a result have less engaged employees compared to leaders who execute a broader and more flexible repertoire of management techniques (Kaiser et al., 2010). They have also found that overdone behavior affects leader effectiveness just as much as underdoing (Kaplan & Kaiser, 2009). Data in support of this is presented in figure 2, where the managers who display “the right amount” of forcefulness also are rated as the most effective leaders.

Figure 2. Relationship between perceived effectiveness and forceful leadership.
(Kaplan & Kaiser, 2009). Either extreme, both underdoing and overdoing the behavior, leads to a marked drop in effectiveness.

In addition to tending towards lopsidedness when over- or underdoing, managers struggle with identifying and judging the degree to which they overdo a behavior. Kaplan and Kaiser (2009) found that 55% of managers were rated by coworkers as using too much of a leadership behavior, but the majority of the leaders did not rate themselves as overdoing that same behavior. This indicates the importance of including overused strengths and lopsided leadership in leadership assessment models, as well as including ratings from several sources.

1.3 Multi-rater assessments

LVI is a multi-rater tool, using the 360-degree method of collecting and comparing feedback from subordinates, peers and superiors, as well as self-assessments. The 360-degree feedback method gets its name from how it includes feedback from different points of view (coworkers on different managerial levels below, on the same level and above the leader in the hierarchy). Multi-rater assessments have become the preferred leader assessment tools in many businesses, as they give leaders a deeper self-understanding (Day, Fleenor, Atwater, Sturm & McKee, 2014). Research shows that 360-tools can enhance communication and performance in organizations (London & Beatty, 1993) and have a positive effect on leader effectiveness (Day et al., 2014).

Because 360-degree assessments use input from different perspectives, the feedback creates a more holistic picture of an individual’s performance and skills compared to traditional top-down, single-source assessments (London & Beatty, 1993; Fleenor & Prince, 1997). Subordinates and peers are often affected by managerial behaviors in ways that are not evident to superiors, and should therefore be included for a valid representation of leader behavior (London & Beatty, 1993). For example, Hooijberg and Choi (2000) found that different rater groups highlight different aspects of the leader role when evaluating effectiveness.

Discrepancies between one’s self-view and the views of others can also give important insight about leadership behavior and effectiveness (Nilsen & Campbell, 1993). Many leaders report that understanding the differences between self-ratings and ratings from others is essential to their development as leaders (Day, 2000).
As multi-rater assessments like the LVI are measures of one leader’s performance rated by several coworkers, the data can be analyzed at two levels: leader level, where each data line represents an assessment of each leader, and rater level, where each data line represents one rater’s evaluation of their target leader. It is important to reflect on which research questions are being explored when choosing level of analysis, as the different levels can shed light on different topics. Leader level data allows for the exploration of individual leader’s behavioral patterns and relationships between different behaviors. Rater level data can inform on the framing of the raters when evaluating the leader, as well as whether the tool is reliable. There will also always be a question of whether the results from one level of analysis are generalizable to the other level.

1.4 The LVI scale
Traditional rating scales used in leader evaluation, including 360-degree feedback instruments, typically use items rated on a Likert-type scale, which represents the degree to which the item characterizes the leader’s behavior. A low rating implies that the leader is doing very little-none of the described behavior, while a high rating indicates that the leader often displays the behavior (Leslie & Fleenor, 1998). An implicit assumption in such a scale is that the higher the rating - or the more often you display a behavior - the better. As assuming that “more is better” leaves little room to explore the possibility of lopsidedness and overdone behaviors, this is a weakness in conventional scales (Kaiser & Overfield, 2010).

To combat the possible occlusion of overdoing behaviors, Kaplan (1996) developed a new rating scale, later revised to the current version by Kaiser and Kaplan (2006). The scale has a range from -4 to +4, where negative scores represent degrees of “too little” and positive scores represent degrees of “too much”, while ratings of 0 represent the appropriate level of the behavior (see figure 3).

![Figure 3. The "too little/too much" rating scale.](image-url)
The idea is that this will give a more nuanced view of the leader’s behavior, as it allows the raters to distinguish between an appropriate amount of a particular behavior and overdone behavior. Research indicates that raters are able to make these distinctions and that the “too little/too much” rating scale is a reliable and valid method for measuring overused strengths (Kaiser & Overfield, 2010). The rating scale also allows for a more nuanced exploration of the theory of lopsided leadership, which proposes that too much of one behavior often results in too little of an opposing behavior. This trend has been found in previous studies of the LVI (e.g. Kaiser & Overfield, 2010).

1.4.1 The versatility measures

In the LVI, the score on the four dimension scales are collected into two total scores, forceful-enabling versatility and strategic-operational versatility, which offers an estimate of the leader’s versatility on each of the how and what domains. The versatility score is an average of the leader’s versatility score on each item pair within the two dimensions the score is based on. The versatility for each pair is computed by utilizing the Pythagorean Theorem, which gives a distance measure from the “right amount” for the leader on those two items (Kaiser & Overfield, 2010; Kaiser et al., 2010).

To have a single score to depict the general performance of leaders, the two versatility factors are made into an overall, average versatility score. The resultant score is a percentage, where higher scores represent a more flexible use of the different leader behaviors - how close the manager is to “the right amount” on all four leadership dimensions. The scores are interpreted similarly to letter grades; A = 90-100 %, B = 80-89 % etc. (Kaiser et al., 2010).

Studies has shown that the versatility score predicts 50 % of the differences of perceived leader effectiveness, meaning that it accounts for half of what separates the most highly regarded leaders from the least well-regarded leaders (Kaiser et al., 2010). Furthermore, versatile leaders have been shown to have employees with higher sense of empowerment, commitment and job satisfaction and their teams have better engagement, cohesion and morale. Being a versatile leader is also associated with producing at a higher quality, quantity and overall output (Kaiser et al, 2010).

1.4.2 Effectiveness indicators

To estimate the performance of leaders, the LVI contains items that estimate three effectiveness indicators: (1) perceived effectiveness of the individual leader, (2) team vitality and (3) team
productivity. Ratings of perceived effectiveness of the leader alone is not necessarily enough to properly gauge a leader’s performance. Therefore, the team vitality and team productivity measures are included, to explore the success of the team the leader is responsible for and thus achieve a more nuanced representation of overall effectiveness (Kaiser, Hogan & Craig, 2008).

Perceived leader effectiveness is measured with ratings on a 10-point scale, where 5 is adequate and 10 is outstanding. Team vitality, concerning the degree of morale, engagement and cohesion in the team, and team productivity, indicating the quantity, quality and overall output of the managers’ teams, are both measured with 5-point scales, where higher scores indicate a more favorable assessment. Research has shown that these scales are valid and that they correlate with similar measures of team productivity and vitality (Kaiser & Overfield, 2010; Kaiser et al, 2010).

Kaiser and Overfield (2010) propose that subordinate ratings are most appropriate for measuring team vitality, while superiors are in the best position to evaluate team productivity. Finally, they suggest that perceived effectiveness should be measured with peer ratings, because peer evaluations have been shown to predict long-term career success (Kraut, 1975). Their research has also supported this segmentation in bettering the prediction of the different effectiveness indicators (Kaiser & Overfield, 2010)

1.5 Differences in American and Norwegian professional culture
To examine the quality of the Norwegian LVI, we will compare the Norwegian data with data from a global sample with a majority of American leaders. When comparing data from different cultures, it is relevant to consider how differences in organizational culture affect the results.

One model through which to understand cultural differences is Hofstede’s cultural dimensions theory. Hofstede has described six dimensions on how countries differ culturally which affect organizational life: power distance, individualism, masculinity, uncertainty avoidance, long-term orientation and indulgence (Hofstede Insights, 2019). Of these, Norwegian culture differs markedly from American (and indeed, most other cultures globally) in the level of masculinity. The femininity-masculinity dimension focuses on how people relate to each other. High levels of femininity are described as “a preference for cooperation, modesty, caring for the weak and quality of life. Society at large is more consensus-oriented.”, while cultures high in masculinity are described as having “a preference in society for achievement, heroism, assertiveness, and material rewards for success. Society at large is more competitive.” (Hofstede Insights, 2019).
Kaiser (2016) have previously found some cultural differences when comparing the 10 most overdone behaviors in the global sample and a sample of Scandinavian leaders. In the global sample, the overwhelming majority were forceful behaviors, while the remainder were operational. In the Scandinavian sample, more operational behaviors and less forceful behaviors were overdone. In addition, one of the top overdone behaviors even fell under the enabling-dimension (Kaiser, 2016). These differences in what behaviors are most often overdone follow a pattern similar to other studies on cultural differences in Scandinavian work culture; it is characterized by a preference for autonomy, independence and equality (Mole, 2003). Furthermore, flat structures, informal communication, empowered employees and little focus on hierarchy often characterize Norwegian businesses (Innovasjon Norge, n.d). Thus, a reduction in the probability of overdoing forceful behaviors in Scandinavia could be a product of cultural restraints.

Exploring the Norwegian translation of the LVI in a Norwegian sample is thus valuable not only due to the opportunity to assess the psychometric qualities of the Norwegian translation; It also offers an interesting arena for generating hypotheses and preliminary findings concerning the cross-cultural validity of the LVI. In the following section, we present our research questions, examining both psychometric qualities of the translation and the exploration of the relationship with relevant effectiveness indicators, as well as possible cultural influences.
2 Rationale and Research Questions

The LVI was translated to Norwegian in 2017, and this is the first psychometric analysis of the LVI on a sample where American culture is not the dominant culture.

2.1 Reliability

2.1.1 Scale reliability

One of the main issues with reliability concerns how much of the variation in the sum of the items can be explained by the common variance created by the underlying construct, or the equivalence of items (Cronbach, 1951). This will be estimated with Cronbach's alpha (α), the most commonly used estimation of scale reliability (Field, 2013). Prior studies show that the four dimension scales (e.g. forceful) and all twelve subscales (e.g. takes charge) exceeds α = .70 in the American version of the LVI 3.0 (Kaiser et al., 2010).

Research question 1: What is the estimated reliability by Cronbach's alpha of the four dimension scales and the twelve subscales in the Norwegian LVI?

2.1.2 Inter-rater reliability

The reliability of the tool also depends on whether ratings are stable from rater to rater, or if the raters are interchangeable. For this multi-rater tool to be reliable, different raters within a rater group must rank leaders similarly, e.g. scoring the same leader in a group as the most versatile, least versatile etc. (LeBreton & Senter, 2008). Previous studies on the LVI have examined inter-rater reliability using Intraclass Correlation Coefficients (ICC) for both single rater [ICC(1)] and for the average rating across multiple raters within each rating group [ICC(k)], where $k$ is the number of raters. Therefore, ICC will be the statistic of choice in this thesis. ICC-values for the American values have previously ranged from .50 to .71 (Kaiser, 2006).

Research question 2: What is the level of inter-rater reliability within rater-groups for the Norwegian LVI?

2.1.3 Inter-rater agreement

Hierarchical models, where lower-level units nest within the higher-order concepts of interest, are a staple in leadership development and organizational psychology research (Cohen, Doveh & Eick, 2001). Because such models utilize averages of single-raters, it is paramount that the raters have a certain level of agreement on scores in order for the averages to be meaningful (Cohen et al., 2001). In previous studies of the LVI, inter-rater agreement has been estimated...
using the James, Demaree and Wolf’s (1984) single-item indices, the \( r_{wg} \)-statistic (Kaiser, 2006). This statistic will therefore be used to allow for comparison. The reported data for the American LVI shows a range from .85 to .93, indicating a high degree of agreement (Kaiser, 2006).

**Research question 3**: What is the inter-rater agreement within rater groups for the Norwegian LVI?

### 2.2 Validity

Historically, there has been much disagreement around how exactly to define validity (see for example; Field, 2013; Thompson, 2004; Goodwin & Leech, 2003). The conceptualization by Cronbach and Meehl (1955) is widely used, which divides validity into four main types: concurrent and predictive validity (which later on was merged into the term criterion-related validity), content validity and construct validity (Cronbach & Meehl, 1955). In recent years, others have challenged this division of validity, arguing that it is more meaningful to see validity as one whole construct, but that there exist several kinds of evidence to support validity. Validity in this conceptualization is seen as an evaluation of whether the intended interpretations are meaningful (Goodwin & Leech, 2003).

While it is quite possible to explore the degree of criterion validity and content validity within the restraint of most modern research, construct validity is much harder to thoroughly evaluate, both due to the abstract nature of constructs and the need for a multimethod-multitrait matrix to ensure both convergent and divergent validity (Campbell & Fiske, 1959). The newer definitions of construct validity also place it as an overarching concept, in which criterion-related and content validity fall within (Goodwin & Leech, 2003). Thus, through evaluating these two forms of validity, one can also be said to be exploring aspects of construct validity. The challenge of criterion-validity is that a proper test of this construct requires certainty that the tool in question is reliable. This is not yet established for the Norwegian LVI. Therefore, although we will explore the relationship between versatility and relevant effectiveness indicators, as well as exploring cross-cultural differences, these analyses are not true tests of criterion validity. Content validity will be explored through several measures, described in more detail below.

#### 2.2.1 Content Validity

A way of exploring content validity is to examine whether the data display patterns in line with the internal structure of the underlying model. According to the theory of lopsided leadership (too much of one behavior often results in too little of an opposing behavior) there should be
negative correlations between the forceful scale and the enabling scale, as well as between the the strategic and operational scales. Measures of Pearson’s correlation coefficient, also referred to as Person’s $r$, of the American LVI show a clear negative relationship between forceful and enabling scales ($r = -.59$), and a weaker negative relationship between the strategic and operational scales ($r = -.21$) (Kaiser and Overfield, 2010).

**Research question 4:** What is the relationship between the forceful and enabling dimensions, and between the strategic and operational dimensions in the Norwegian LVI?

Another way to explore the internal structure is whether the items display the expected negative correlations with their opposing item in the corresponding dimension. We did not find any data on the item correlations for the global sample. Therefore, these results will not be compared to global trends.

**Research question 5:** What is the relationship between the opposing items in the dimension pairs forceful-enabling and strategic-operational?

The third aspect of the internal structure of the LVI that warrants exploration is the fit of the suggested model. In previous studies of the American LVI, the validity of the theoretical model has been examined using exploratory and confirmatory factor analysis, to explore whether the data recreated the structure expected from the theoretical model. Using maximum likelihood factor analysis, they found that the patterns of factor coefficients clearly represent the forceful, enabling, strategic and operational dimensions. Furthermore, forceful and enabling were more strongly associated with one another than strategic-operational and vice versa. The four dimensions were estimated to account for 58.6% of the common variance after rotation (Kaiser, 2006). Finally, the confirmatory factor analysis on global data found satisfactory model fit for the forceful-enabling dimension pair in the American LVI. The model fit for the strategic-operational pair has not been explored in the global sample, due to the version of the scale available at the time not conforming empirically to the model (Kaiser, 2006).

In the global samples, the model has been analyzed on leader level with the aggregated scores of the different leaders. As we only have enough data points for such analyses if we utilize the rater level data, the subsequent analyses cannot say anything about whether the theoretical model fits well with the behavioral patterns of leaders. What an analysis on rater level data can explore, is the frame of mind the individual raters have when evaluating the leaders. In other words, the
following analyses explore how raters use the Norwegian LVI tool and whether this follows the patterns predicted by the theoretical model.

*Research question 6:* What is the internal structure of the model on rater level data in the Norwegian LVI?

### 2.2.2 Relationship with effectiveness indicators

Although there is a question of whether the scale is reliable, how the versatility scores relate to relevant variables such as effectiveness measures can still serve as an early exploration of the concurrent validity, as an aspect of criterion validity. Previous studies on the LVI have found a curvilinear relationship between the four leadership dimensions and perceived leader effectiveness, where the highest degree of effectiveness relates to “the right amount” of the specific behavior, and that both overdone and underdone behavior affects leader effectiveness negatively (Kaiser, 2006). The relationship thus explores whether overdoing negatively affects leadership effectiveness and whether versatility correlates significantly with perceived leadership effectiveness.

*Research question 7:* What is the relationship between the four dimensions (forceful, enabling, strategic and operational) and perceived leader effectiveness in the Norwegian LVI?

American research has shown that both forceful-enabling versatility and strategic-operational versatility are good predictors of ratings of leader effectiveness: About half of the variance in effectiveness ratings in global samples has been found to be a function of leader versatility within the how and what domains (Kaiser, 2006). How effectiveness relates to versatility will therefore be explored in the Norwegian LVI as well.

*Research question 8:* What is the relationship between the two versatility measures (forceful-enabling and strategic-operational versatility) and perceived leader effectiveness in the Norwegian sample?

Finally, the LVI contains measures of team productivity and team vitality. Previous studies have found that forceful-enabling versatility most strongly predicts team vitality as rated by subordinates, while strategic-operational versatility was found to be the best predictor of team productivity as rated by superiors. This supports the conceptualization of forceful-enabling versatility as the how of leadership, and strategic-operational versatility as the what of
leadership. Overall perceived effectiveness has previously been measured by peer ratings in the global sample analyses. This will therefore also be done in the Norwegian sample.

Research question 9: When segmenting the effectiveness indicators into subgroup-ratings, how does forceful-enabling and strategic-operational versatility relate to perceived leader effectiveness, team vitality and team productivity?

2.2.3 Differences in frequency of leadership behaviors

Another external variable that may shed light on the qualities of the Norwegian LVI is whether there are any cultural differences in frequency, distribution or relationships between variables in the Norwegian versus the global data, as well as whether these match previously established differences in organizational culture.

Research question 10: Are there any clear differences between the results from the Norwegian and global sample, and do they coincide with previously established cultural differences in organizational culture?
3 Method

3.1 Administration

The LVI contains 55 items and takes approximately 10 to 15 minutes to complete. It is designed to collect data from key coworkers, categorized as superiors, peers and subordinates, and from the target leader herself or himself. To protect the confidentiality of rater responses, there is a lower limit of three raters for the rater groups at the lower or same hierarchical level as the leader.

3.2 Sample

A quantitative, non-random design with convenience sampling was used in this study. The sample involved ratings of 44 managers, with 543 coworker ratings (212 peer ratings, 290 subordinates and 41 superiors), using the Leadership Versatility Index multi-rater assessment instrument (Kaiser et al., 2010). The majority of the raters were from governmental Norwegian firms, with about 34 % working in the private sector. On average, managers were rated by 12 coworkers. The mean age of the rated managers was 51 years, ranging from 29 to 64, with a balanced gender ratio of 49 % males and 51 % females. Most of the respondents completed the survey using the Norwegian translation, but 35 did it in English. These responses were removed from the data set before the analyses, as the goal of the study was to assess the psychometric qualities of the Norwegian version of the LVI. When performing the analyses of the data, self-ratings were excluded, as previous studies have found a low degree of overlap between how leaders score themselves and how their coworkers’ rate them (Kaplan & Kaiser, 2006; Kaplan & Kaiser, 2009).

The data were compared to data from several different studies, based on different samples. All samples were of mainly North American background, with a majority of male leaders. The global sample, which the descriptive data and data on rates of overdoing and underdoing was based on, consisted of 18 216 leaders. The majority of the leaders were from the United States (46 %), while the rest of the sample was 28 % European, 11 % Asian, 5 % Latin-American, 5 % African/Middle-Eastern and 4 % Australian. In addition, the sample was 69 % male and the mean age was 42 (Kaiser Leadership Solutions, 2016). We could not find detailed demographic information on the other samples. Because the global sample varies for different analyses, the sample size of the global sample will be specified in every analysis.
3.3 **Ethical and legal considerations**

The data collection and storage were approved by the NSD - Norwegian Centre for Research Data. As the raters were anonymous, the leaders anonymized prior to analysis and all research participants were typically functioning adults, the study was not thought to have any special ethical considerations as to the well-being of the participants.

3.4 **The Norwegian translation**

The ratings in our sample were collected with the Norwegian translation of *Leadership Versatility Index* version 4.0. The translation was a collaboration between the students and authors of this thesis (Ida Kopperstad and Sofie Bang Jensen), associate professor Henning Bang and specialist in organizational psychology Thomas Nesset Midelfart. A professional translator translated the final translation back to English, to ensure that the core meaning of the questions had not been lost. An example of a statement from the Norwegian translation can be seen in figure 3 (p. 9).

3.5 **Level of analysis**

For the reliability analyses, all rater scores were used, as reliability is not related to the specific leader. The same level is used for the factor analyses due to issues of sample size. For predictive validity, on the other hand, the hypothesis is linked to whether a leader's LVI-score could predict their performance on relevant effectiveness indicators. Therefore, the data have to be analyzed on leader level, resulting in a much smaller sample of 44 managers, rather than the 543 data lines of the rater level data.
4 Results

4.1 Descriptive statistics

To explore the qualities of the distribution, descriptive statistics were undertaken for the versatility scores, effectiveness indicators and the four dimensions, presented in table 1, 2 and 3, respectively. The Norwegian sample was analyzed on leader level data, with a total of 44 managers. The global results were based on coworker ratings for 18,216 target managers (Kaiser Leadership Solutions, 2016).

Table 1
Sample size, range, mean and standard deviation for the versatility scores in the Norwegian and global sample.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Norwegian sample</th>
<th>Global sample</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>Range</td>
</tr>
<tr>
<td>F-E versatility</td>
<td>44</td>
<td>73 % - 97 %</td>
</tr>
<tr>
<td>S-O versatility</td>
<td>44</td>
<td>69 % - 97 %</td>
</tr>
<tr>
<td>Overall versatility</td>
<td>44</td>
<td>71 % - 97 %</td>
</tr>
</tbody>
</table>

The Norwegian sample was clearly affected by restriction of range in the lower end of the spectrum. In accordance with this, the mean scores were high, ranging from 87 % to 88 %. The global sample had a broader range, but the mean scores were still high, between 82 % and 85 %.
Table 2
Sample size, range, mean and standard deviation for the effectiveness indicators in the Norwegian and global sample.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Norwegian sample</th>
<th>Global sample</th>
<th>Norwegian sample</th>
<th>Global sample</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>Range</td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>Perceived leader effect.</td>
<td>44</td>
<td>6.18 - 8.95</td>
<td>7.78</td>
<td>.66</td>
</tr>
<tr>
<td>Team vitality</td>
<td>44</td>
<td>1.63 - 3.07</td>
<td>3.87</td>
<td>.42</td>
</tr>
<tr>
<td>Team prod.</td>
<td>44</td>
<td>1.30 - 3.12</td>
<td>3.74</td>
<td>.38</td>
</tr>
</tbody>
</table>

The trend of a restricted lower range in the Norwegian sample compared to the global results was also evident for the effectiveness indicators. It was especially prominent for the lower spectrum on the perceived leader effectiveness scale. However, the mean scores of the Norwegian and global samples were quite similar.

Table 3
Sample size, range, mean and standard deviation for the four dimensions in the Norwegian and global sample.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Norwegian sample</th>
<th>Global sample</th>
<th>Norwegian sample</th>
<th>Global sample</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>Range</td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>Forceful</td>
<td>44</td>
<td>-.80 -.38</td>
<td>-.20</td>
<td>.27</td>
</tr>
<tr>
<td>Enabling</td>
<td>44</td>
<td>-.53 -.23</td>
<td>-.14</td>
<td>.19</td>
</tr>
<tr>
<td>Strategic</td>
<td>44</td>
<td>-1.39 -.16</td>
<td>-.32</td>
<td>.31</td>
</tr>
<tr>
<td>Operational</td>
<td>44</td>
<td>-.40 -.25</td>
<td>-.06</td>
<td>.15</td>
</tr>
</tbody>
</table>

The restriction of range was also present for the four dimensions. The highest score in the Norwegian sample was .38 for forceful leadership, while the lowest was -1.39 for strategic leadership. Again, the global sample had a much broader range, as expected with a larger sample.
The frequency of versatile leaders was also explored, to gauge whether there were any differences in the versatility of the leaders in the Norwegian sample compared to the global sample. The results are presented in table 4 and figure 4.

Table 4  
*Overview of the distribution of versatility scores in the Norwegian and global sample*

<table>
<thead>
<tr>
<th>Versatility score</th>
<th>Norwegian sample (N = 44)</th>
<th>Global Sample (N = 18,216)</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 70 %</td>
<td>0 %</td>
<td>4 %</td>
</tr>
<tr>
<td>70-80 %</td>
<td>7 %</td>
<td>19 %</td>
</tr>
<tr>
<td>80-90 %</td>
<td>59 %</td>
<td>68 %</td>
</tr>
<tr>
<td>90-100 %</td>
<td>34 %</td>
<td>9 %</td>
</tr>
</tbody>
</table>

*Figure 4. A visual representation of the Norwegian sample distribution on overall versatility on the left, while the global sample is presented on the right.*

As can be seen from figure 4, both samples were negatively skewed; the majority of leaders scored 80 % or above. Although there was a degree of kurtosis and skewness in the sample, it is generally accepted that kurtosis and skewness does not interfere unduly with results of parametric analyses (Gravetter & Wallnau, 2014). Previous studies have found parametric tests to perform well even for non-normal distributions, as long as the variables are independent (e.g. Edgell & Noon, 1984; Rasmussen, 1987). Therefore, the use of parametric tests was seen as justified. Although the distributions were similar, the Norwegian sample had an even higher frequency of leaders who had an overall versatility score above 90 %.
4.2 Reliability

4.2.1 Scale reliability

To estimate the scales’ reliabilities in the Norwegian sample, we computed Cronbach’s alpha for the four dimensions and 12 subscales. The alpha values for each dimension and subscale for the Norwegian and global sample are presented in table 5. The global data are based on ratings from 12 620 raters (Kaiser et al., 2010), while the Norwegian analysis is conducted on rater level data of 543 raters.

Table 5
Reliability scores for the four dimensions and 12 subscales in the Norwegian and global sample

<table>
<thead>
<tr>
<th>LVI Dimension (12 items)</th>
<th>Norwegian sample version 4.0 (N = 543)</th>
<th>Global sample version 3.0 (N = 12 620)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>α</td>
<td>α</td>
</tr>
<tr>
<td>Forceful</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F1 Takes charge</td>
<td>.86</td>
<td>.89</td>
</tr>
<tr>
<td>F2 Declares</td>
<td>.70</td>
<td>.78</td>
</tr>
<tr>
<td>F3 Pushes</td>
<td>.81</td>
<td>.77</td>
</tr>
<tr>
<td></td>
<td>.75</td>
<td>.81</td>
</tr>
<tr>
<td>Enabling</td>
<td>.84</td>
<td>.90</td>
</tr>
<tr>
<td>E11 Empowers</td>
<td>.84</td>
<td>.85</td>
</tr>
<tr>
<td>E2 Listens</td>
<td>.84</td>
<td>.81</td>
</tr>
<tr>
<td>E3 Supports</td>
<td>.61</td>
<td>.84</td>
</tr>
<tr>
<td>Strategic</td>
<td>.93</td>
<td>.92</td>
</tr>
<tr>
<td>S1 Direction</td>
<td>.91</td>
<td>.86</td>
</tr>
<tr>
<td>S2 Growth</td>
<td>.84</td>
<td>.83</td>
</tr>
<tr>
<td>S3 Innovation</td>
<td>.81</td>
<td>.80</td>
</tr>
<tr>
<td>Operational</td>
<td>.64</td>
<td>.81</td>
</tr>
<tr>
<td>O1 Execution</td>
<td>.70</td>
<td>.78</td>
</tr>
<tr>
<td>O2 Efficiency</td>
<td>.52</td>
<td>.70</td>
</tr>
<tr>
<td>O3 Order</td>
<td>.44</td>
<td>.71</td>
</tr>
</tbody>
</table>

A cut-off of .70 usually attributed to Nunally (1978) is widely accepted as the standard for alpha values (Lance, Butts & Michels, 2006). The actual statement is much more nuanced in its view on cut-offs, establishing that one cut-off should not be applied universally (Nunnally, 1978). Nevertheless, as previous studies on the American LVI have used the cut-off of .70 (e.g. Kaiser, 2006; Kaiser et al., 2010) and that it is a commonly accepted value in the research literature in
general (see for example Kline, 2000; DeVellis 2012), it was seen as an appropriate cut-off for this thesis.

Three of the four dimensions - forceful, enabling and strategic - exceeded the preferable .70 alpha value. The operational dimension had a lower internal consistency, at .64. The subscales nested under the forceful and strategic scales all exceeded .70. Under enabling, the subscale supports had a lower alpha value (.61) compared to the other subscales (.84). This difference was not found in the global sample. The alpha values of the subscales under the operational dimension ranged from .44 to .70 in the Norwegian sample. This also differs from the global sample, where all alpha values were estimated at .70 and above.

To examine the operational scale in more detail, the inter-item correlation for the items in this dimension as well as alpha if item deleted were analyzed. These values are presented in appendix 1. There were multiple items with zero-value correlations and even some that correlated negatively. The items 8 (in the subscale o2 efficiency), 11 and 12 (both in the subscale o3 order) showed increased alpha values for their respective subscales if the item was deleted.

Cronbach’s alpha assumes that the scale is internally consistent - that the items of the scale are measuring the same underlying attribute (Throchim, 2006b). If this is not the case, it can affect the alpha score negatively. To assess possible multidimensionality, the internal consistency of the scales was explored, by extracting components with eigenvalue greater than 1 using principal component analysis (PCA). PCA is a technique for identifying clusters of variables and understand the structure of a set of variables (Field, 2013). As can be seen in appendix 2, o2 efficiency and o3 order in the operational scale were multidimensional. Except for the subscale s3 supports within the enabling scale, none of the other dimensions had any multidimensional subscales.

4.2.2 Inter-rater reliability

The reliability of a tool that uses averages of raters is dependent on a certain degree of stability of one rated leader’s score across different raters. Estimates of inter-rater reliability indicate the rank-order consistency of leader rankings by raters, i.e. the consistency of multiple raters rating the different targets (LeBreton, Burgess, Kaiser, Atchley & James, 2003). To examine the inter-rater reliability of the scale, Intraclass Correlation Coefficients (ICC) were computed for average measures on all four dimensions. Only ICC-statistics for the peer and subordinate-groups were computed in the Norwegian sample, as all leaders were evaluated by one superior
each. Based on recommendations from Landers (2015), a one-way random ICC was computed, as the subjects are rated by different raters and the raters are selected from a larger population of possible raters. Average measures were chosen due to the fact that the scales are expected to be used by minimum three people for both the peer and subordinate groups in real life application. Values above .70 is the preferred cutoff for ICC($k$), but scores above .50 are usually accepted (Koo & Li, 2016).

The results are presented in table 6. The data was analyzed on rater level with 543 raters in the Norwegian sample, while the global data was based on average values calculated across six independent samples with 4 878 raters (Kaiser et al., 2010). In the Norwegian sample, $k$ was equal to two. The global samples had $k$-values at four for both groups (Kaiser, 2010).

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Rater group</th>
<th>Norwegian sample ($N = 543$)</th>
<th>ICC(2)</th>
<th>Global sample ($N = 4 878$)</th>
<th>ICC (4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Forceful</td>
<td>Peers</td>
<td>.75</td>
<td></td>
<td>.71</td>
<td></td>
</tr>
<tr>
<td>Enabling</td>
<td>Peers</td>
<td>.61</td>
<td></td>
<td>.70</td>
<td></td>
</tr>
<tr>
<td>Strategic</td>
<td>Peers</td>
<td>.59</td>
<td></td>
<td>.53</td>
<td></td>
</tr>
<tr>
<td>Operational</td>
<td>Peers</td>
<td>.70</td>
<td></td>
<td>.54</td>
<td></td>
</tr>
<tr>
<td>Forceful</td>
<td>Subordinates</td>
<td>.68</td>
<td></td>
<td>.64</td>
<td></td>
</tr>
<tr>
<td>Enabling</td>
<td>Subordinates</td>
<td>.56</td>
<td></td>
<td>.58</td>
<td></td>
</tr>
<tr>
<td>Strategic</td>
<td>Subordinates</td>
<td>.76</td>
<td></td>
<td>.53</td>
<td></td>
</tr>
<tr>
<td>Operational</td>
<td>Subordinates</td>
<td>.58</td>
<td></td>
<td>.50</td>
<td></td>
</tr>
</tbody>
</table>

All scores in the Norwegian sample exceeded .50. Several scores also achieved the level of .70, thereby qualifying as “good” (Koo & Li, 2016). In addition, the Norwegian ICC-values were in general equal to or somewhat higher than the global values, with the exception of the enabling ratings from peers. The peer ratings of the operational scale and subordinate ratings of the strategic scale in the Norwegian sample were markedly higher than the corresponding values in the global sample.
4.2.3 Inter-rater agreement

For the LVI, as in other multilevel models, much of the research on validity is done with average scores within the different rater groups. This is done to access a broader construct, “as perceived by this rater-group”. Feedback to leaders is also based on the average scores of the different rater groups, to ensure confidentiality. Therefore, it is essential to explore the agreement within rater groups, to justify aggregation of lower level-items and in representing group scores as averages (Cohen et al., 2001). Inter-rater agreement refers to the degree the different scores are interchangeable. It is an estimate of “the absolute consensus in scores furnished by multiple judges for one or more targets” (LeBreton & Senter, 2008, p. 816). To evaluate this assumption in the sample, the $r_{wg}$ statistic was used. The statistic indicates the proportion of observed variance that could be a result of random error.

The choice to evaluate interrater agreement in addition to interrater reliability is in part because the two estimates are differentially vulnerable to biases. As measures of inter-rater reliability such as Pearson’s r or ICC are vulnerable to the effects of restriction of range in sample variance, the $r_{wg}$ statistic may give a more realistic picture of inter-rater similarity between groups in 360-degree tools (LeBreton et al, 2003). Furthermore, inter-rater reliability is typically estimated through a type of correlation coefficient, while inter-rater agreement is estimated by within-group dispersion. Thus, the $r_{wg}$-statistic is for the most part unaffected by number of raters, while the ICC-score will increase with the number of raters (LeBreton & Senter, 2008).

Mean $r_{wg}$-values greater than .70 are generally regarded as good, with values closer to 1.00 indicating a very high degree of agreement (Bliese, 2000; James et al., 1984). A .70 score entails that 30 % of variance in scores should be contributed to error. LeBreton and Senter (2008) suggest that a less rigid cut-off might be a better choice, where the researcher can evaluate the necessary level of agreement. They categorize .00 to .30 as “lack of agreement”, .31 to .50 as “weak agreement”, .51 to .70 as “moderate agreement”, and finally .71 to .90 as “strong agreement” and .91 to 1.00 as “very strong agreement”.

A common critique of the $r_{wg}$ statistic is the assumption of a uniform null distribution, or a rectangular distribution of scores. Since human beings are prone to a variety of response biases (leniency bias etc.), this will not always be the case (LeBreton & Senter, 2008). Nevertheless, as the uniform distribution has been the most used and that there are arguments in favor of it as well as against (LeBreton & Senter, 2008), it is seen as suitable for the present study. The results
of the $r_{wg}$-analyses on rater level data are shown in table 7, together with the results from the global sample (Kaiser et al., 2010).

Table 7
$r_{wg}$-analyses for the four dimension scales, as rated by peers and subordinates in the Norwegian and global sample

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Rater group</th>
<th>Norwegian sample $(N = 543)$ $r_{wg}$</th>
<th>Global sample $(N = 4,878)$ $r_{wg}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Forceful</td>
<td>Peers</td>
<td>.82</td>
<td>.85</td>
</tr>
<tr>
<td>Enabling</td>
<td>Peers</td>
<td>.87</td>
<td>.88</td>
</tr>
<tr>
<td>Strategic</td>
<td>Peers</td>
<td>.84</td>
<td>.89</td>
</tr>
<tr>
<td>Operational</td>
<td>Peers</td>
<td>.87</td>
<td>.88</td>
</tr>
<tr>
<td>Forceful</td>
<td>Subordinates</td>
<td>.87</td>
<td>.85</td>
</tr>
<tr>
<td>Enabling</td>
<td>Subordinates</td>
<td>.90</td>
<td>.89</td>
</tr>
<tr>
<td>Strategic</td>
<td>Subordinates</td>
<td>.88</td>
<td>.89</td>
</tr>
<tr>
<td>Operational</td>
<td>Subordinates</td>
<td>.59</td>
<td>.86</td>
</tr>
</tbody>
</table>

The results of the Norwegian sample were for the most part quite similar to the global sample, except for the operational dimension as rated by subordinates. The forceful, enabling and strategic scales showed similar values for both subordinate and peer rating groups. All the $r_{wg}$-values of the three scales were also well above .70, which indicates a high degree of interrater agreement (LeBreton & Senter, 2008). The one exception to this trend was the $r_{wg}$-value on the operational dimension scale as scored by the subordinate rater group. The score of .59 indicates a moderate degree of interrater agreement.

4.3 Validity

4.3.1 Content validity - Relationship between opposites

As described earlier, the underlying theory of the LVI assumes that managers tend to overdo one type of leadership behavior at the expense of another, referred to as lopsidedness (Kaplan & Kaiser, 2009). Therefore, a negative correlation between forceful and enabling leadership scores, and between strategic and operational leadership would support the content validity of the Norwegian LVI. Such patterns have previously been found for the American LVI in a global
sample (based on coworker ratings of 484 leaders), with Pearson’s r at -.59 and -.21 for the forceful-enabling and strategic-operational dimension pairs, respectively (Kaiser & Overfield, 2010). Figure 5 presents scatter plots illustrating the relationship between the dimension pairs for the Norwegian sample, analyzed on leader level data (N = 44).

![Figure 5. Relationship between the opposing behaviors.](image)

A significant negative correlation between the forceful and enabling scale was found, with r-value -.60 (p ≤ .01). Overdone behavior at the forceful scale was related to underdone behavior at the enabling scale and vice versa. The relationship between the strategic and operational scales was weaker, with a non-significant, negative correlation of -.27. These results follow the same pattern as the global sample, where the strategic and operational dimensions display a less clearly opposite relationship than the forceful-enabling dimension pair (Kaiser & Overfield, 2010).

The pattern of opposites should also be found on item- and subscale-level, as the versatility scores are computed on item level (Kaiser, 2006). To explore the internal structure of the versatility measures, the correlations between opposite items and opposite subscales were therefore computed. The results are presented in appendix 3 (item-level) and 4 (subscale-level). Most of the opposite item-pairs had a correlation ranging from around 0 to -.40. One item pair within the s1 direction - o1 execution subscales had a positive correlation of .39. There was therefore a substantial variation in how the item pairs conformed to the theoretical assumptions of the model. The correlations between the opposing subscales were a somewhat better fit with the model of opposites, as all opposing pairs except s1 direction - o1 execution and s3 innovation - o3 order displayed negative correlations above -.28. Again, s1 direction - o1 execution was the weakest link, with a correlation around zero.
4.3.2 Content validity - Exploration of the conceptual structure

The second aspect of content validity for the Norwegian LVI concerns how well the data fit the conceptual structure suggested by the theoretical model. This has previously been explored using both exploratory and confirmatory factor analysis (Kaiser, 2006; Kaiser et al., 2010). As described earlier, it was not feasible to explore the patterns in leader level data due to the small sample. Instead, we explore rater level data to analyze how raters use the tool and whether the system through which they evaluate the leaders follow the trends expected from the theoretical model.

Principal Component Analysis

To explore whether the underlying model of the LVI is supported by patterns in the data set, principal component analysis (PCA) with varimax rotation was used. PCA aims to explain the maximum amount of total variance by transforming the original variables into linear components. This is in contrast to factor analysis, which attempts to achieve parsimony by explaining the maximum amount of common variance (Field, 2013). Furthermore, PCA also allows for a more thorough exploration of all possible loadings and complex items, while only the items specified as complex are allowed to load on several factors in confirmatory factor analysis (Thompson, 2004). As PCA is focused on establishing which linear components exist within the data and how a particular variable might contribute to that component (Field, 2013), it is seen as the appropriate method for the purpose of this exploratory analysis. The use of PCA with varimax rotation is well established in validation studies within several fields of psychology and psychometric evaluation (Laher, 2010).

Earlier studies analyzing the LVI have utilized oblimin rotation for these analyses. Our choice to use varimax rotation was due to oblimin rotation possibly reducing the value an exploratory factor analysis adds to the results. When the factors are allowed to correlate as in oblimin rotation, correlations can both be a result of the correlated factors and that subscales load on several factors, so-called complex items (Lei & Wu, 2007). In varimax rotation, the factors are not allowed to correlate, which means that correlations are exclusively the result of complex items (Lei & Wu, 2007), thus adding insight and value beyond the data from the confirmatory factor analysis, where only the relationships designated with the model are included (Timothy, 2010).
We conducted the analysis with the 12 subscales on rater level data. The number of factors selected was based on eigenvalue, with values of one as the cut-off value, in line with Kaiser’s criterion (Kaiser, 1960). It is based on the fact that eigenvalues represent the amount of variation explained by a factor. An eigenvalue greater than one indicates that the component explains more variation than an average single item, which is seen as a substantial amount of variation (Field, 2013).

When extracting the components with eigenvalue higher than one, three components was retained, in contrast to the four extracted in the global sample (Kaiser et al., 2010), as shown in table 8.

Table 8
Rotated component matrix, PCA without predetermined components

<table>
<thead>
<tr>
<th>Component</th>
<th>Component 1</th>
<th>Component 2</th>
<th>Component 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1 Takes charge</td>
<td>.78</td>
<td>.31</td>
<td>.18</td>
</tr>
<tr>
<td>F2 Declares</td>
<td>.73</td>
<td>.15</td>
<td>-.13</td>
</tr>
<tr>
<td>F3 Pushes</td>
<td>.65</td>
<td>.32</td>
<td>-.04</td>
</tr>
<tr>
<td>E1 Empowers</td>
<td>-.70</td>
<td>.18</td>
<td>-.26</td>
</tr>
<tr>
<td>E2 Listens</td>
<td>-.50</td>
<td>.63</td>
<td>.32</td>
</tr>
<tr>
<td>E3 Supports</td>
<td>-.58</td>
<td>.34</td>
<td>.44</td>
</tr>
<tr>
<td>S1 Direction</td>
<td>.25</td>
<td>.81</td>
<td>-.05</td>
</tr>
<tr>
<td>S2 Growth</td>
<td>.29</td>
<td>.83</td>
<td>-.26</td>
</tr>
<tr>
<td>S3 Innovation</td>
<td>.03</td>
<td>.85</td>
<td>-.19</td>
</tr>
<tr>
<td>O1 Execution</td>
<td>.49</td>
<td>.17</td>
<td>.53</td>
</tr>
<tr>
<td>O2 Efficiency</td>
<td>-.02</td>
<td>-.24</td>
<td>.69</td>
</tr>
<tr>
<td>O3 Order</td>
<td>.02</td>
<td>-.12</td>
<td>.70</td>
</tr>
</tbody>
</table>

In the component matrix the forceful and enabling subscales appeared to load on the same component, but with opposite component loadings. The strategic and operational subscales, however, loaded on two different components. In addition, there were some subscales that loaded on several components; o1 execution, e2 listens and e3 supports. The three components
in the rotated solution explained 25 %, 24.4 % and 14.7 % of the variance, respectively, resulting in a total variance explained at 64.1 %.

An alternative interpretation of the LVI-structure is a two-dimensional model, where *how* and *what* represents the two dimensions, while the behavioral categories (forceful, enabling, strategic and operational) represent the opposing poles on the associated dimension. As the expected four-component was not replicated in the exploratory factor analysis, we decided to explore a model based on this interpretation as well, to see whether it was a better fit to the data. This also allows for an exploration of whether there are signs of over-factorization in the model. We conducted a PCA with two predetermined components to investigate how the subscales behaved in such a model. The rotated component plot is presented in figure 6. The associated component matrix can be found in appendix 5.

The component plot shows that the subscales clustered together in four groups. This was especially evident for the enabling subscales (e1-e3), the strategic subscales (s1-s3) and for the forceful subscales (f1-f3). The operational scale, on the other hand, showed a higher degree of diversification, where subscale o1 *execution*, was clustered closer to the components from the
forceful scale. All subscales in the operational scale (o2 efficiency and o3 orders) were also located considerably closer to origo than the other subscales.

The rotation of the sum square loadings shows that the what-component explained 25.53% and the how-component 25.03 % of the variance in the sample, resulting in a total explained variance at 50.56 %. In other words, half of the variance in the data can be explained by a two-dimensional solution, approximately 14 % less than the three-factor solution.

**Confirmatory Factor Analysis**

Confirmatory factor analysis (CFA) differs from exploratory factor analysis in that the desired number of factors and the relationships between them is specified and superimposed on the data set before the analysis. This means that the criteria for the analysis is the degree of fit of the hypothesized model to the tendencies observed in the data, where the goal is to minimize the difference between the hypothesized and actual covariance matrices (Schreiber, Nora, Stage, Barlow & King, 2006).

The confirmatory factor analysis in this study was performed using Structural Equation Modeling (SEM; Schreiber et al., 2006). SEM is a commonly used method and can be seen as a combination of the principles behind confirmatory factor analysis and path models, as it utilizes both observed and latent variables, thus allowing for the modeling and exploration of complex phenomenon (Schumacker & Lomax, 2004, p. 5; Timothy, 2010). It also explicitly describes measurement error in the model, which is an advantage (Raykov & Marcoulides, 2006).

SEM-modeling does have stringent assumptions to the sample; there must be a reasonable sample size, the variables must be continuously and normally distributed and there must be a minimum of known correlations as input (Schreiber et al., 2006). In the following analysis, rater level data with 543 raters was used, within the desired minimum of 200 raters suggested by e.g. Kline (2005). Others recommend at least 10 raters per estimated parameter (Schreiber et al., 2006), which is also achieved in this model. Nevertheless, there is still a question of whether we have a sufficient number of raters for the analysis due to the sample not being normally distributed, and the results must be viewed in light of this. It was still considered worthwhile to attempt the analysis as SEM has a unique ability to evaluate the fit of whole models (Schreiber et al., 2006). Maximum likelihood estimation was used.

The model explored in the CFA was the original LVI model, with four separate dimensions. The analyses performed on the global sample data utilized single items as observed variables.
(Kaiser, 2006). Due to the small sample size, this was not possible in our thesis. Therefore, the model explored focused on a higher level of the hierarchical model, using the subscales as the observed or manifest variables. The decision was also based on the higher level of reliability achieved with parceling compared to using single items. Previous studies on the LVI-model have only performed a CFA for the forceful-enabling pair (Kaiser, 2006). The forceful-enabling and strategic-operational pair were therefore tested in separate models, to allow for better comparison with the American data. The two dimensions within each dimension pair were given a bidirectional relationship, while the subscales were defined as having one-way, causal relationships with their respective dimension.

There exist many different measures of fit in SEM. The chi-square is a widely known statistic that assesses overall fit of the model and will therefore be included, although the statistic is sensitive to sample size and non-normal distributions (Kenny, 2015), which is not ideal in our sample. In addition, we will examine the Tucker-Lewis Index (TLI), which is considered better for smaller samples (Kline, 2005). The recommended cut-off value from Schreiber et al (2006) is .95 for TLI. The Root Mean Square Error of Approximation (RMSEA) will also be used, as it is the most common assessment of model fit (Kenny, 2015). Some set the recommended cut-off value below .06, but it is also commonly accepted that values up to 1 are acceptable (Kenny, 2015). Schreiber and colleagues (2006) also recommend both these indices. The indices were compared to the values achieved in the American version (Kaiser, 2006), although the level of analysis was different. The fit indices are presented in table 9, while the models for the Norwegian data are presented visually in figure 7. The visual representation of fit for the LVI model in the global sample can be found in appendix 6.
### Table 9

*Fit indices for the confirmatory factor analysis*

<table>
<thead>
<tr>
<th>Fit indices</th>
<th>Norwegian sample FE</th>
<th>Norwegian sample SO</th>
<th>Global sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>543</td>
<td>543</td>
<td>2839</td>
</tr>
<tr>
<td>$x^2$</td>
<td>195.56*</td>
<td>58.47*</td>
<td>1962.55</td>
</tr>
<tr>
<td>df</td>
<td>8</td>
<td>8</td>
<td>109</td>
</tr>
<tr>
<td>TLI</td>
<td>.46</td>
<td>.87</td>
<td>.89</td>
</tr>
<tr>
<td>RMSEA</td>
<td>.20 (.18 - .23)</td>
<td>.10 (.80 - .13)</td>
<td>.07</td>
</tr>
</tbody>
</table>

*significant model chi-square value

As can be seen from the table, the scores on the fit indices were lower for the Norwegian sample compared to the global sample. Both chi-squares were, as expected, significant. The TLI- and RMSEA-values did not achieve cut-off for both models, although the strategic-operational model was somewhat closer to the cut-off values for both measures. The forceful and enabling factors correlated at approximately the same level as in the model for the global sample (-.54 compared to the global -.58), while the strategic and operational factors were not correlated (-.02).

**Figure 7.** Illustration of the CFA-models for the Norwegian sample, with standardized estimates.
4.4 The Norwegian LVI and effectiveness indicators

4.4.1 The relationship between perceived leader effectiveness and the four dimensions

An important assumption of the LVI is that both underdone and overdone behavior affects leader effectiveness negatively (Kaplan & Kaiser, 2009). This implies a curvilinear relationship between how managers are rated on the four dimensions, and to what extent they are rated as effective leaders. Figure 8 displays these relationships in the Norwegian sample for the four dimensions, including the corresponding curvilinear regression line, analyzed on leader level (N = 44).

As can be seen from the figures above, there was a weak, curvilinear relationship between the four dimensions and perceived leader effectiveness. Curvilinear regression lines had a better fit compared to linear regression lines, but the explained variance was still small; the relationships ranged from a non-significant .04 and .08 for the operational and the enabling scale, to a significant .14 (p ≤ .05) and .25 (p ≤ .01) for the forceful and the strategic scale, respectively.

Figure 8. The relationship between perceived leader effectiveness and the four dimensions.
Ratings on leader effectiveness below 6.5 are usually defined as relatively low, and scores above 8.5 as relatively high (Kaiser et al., 2010). The highly rated leaders fell within +/- .40 on the scale for forcefulness, -.60 to +.20 on the strategic scale, and from -.40 to +.20 on both the enabling and operational scale. There were also leaders that were rated as less effective within these spectrums. The majority of the leaders was rated between 7 to 9 on effectiveness with no leaders rated below 6, resulting in a restriction of range on the effectiveness scale. There was also a restriction of range in the leader behavior scales, especially for the operational scale, where all leaders were rated between -.40 and +.30.

4.4.2 The relationship between perceived leader effectiveness and versatility

Another way to investigate the relationship between the four dimensions and perceived leader effectiveness in the Norwegian LVI, is to examine the correlation between perceived effectiveness and the different versatility scores (forceful-enabling, strategic-operational and overall versatility) with Pearson’s r. The results for the Norwegian sample analyzed on leader level are presented in table 10. The global numbers are based on average coworker ratings of 450 target leaders (Kaiser, 2006).

Table 10
Correlations between perceived leader effectiveness and versatility in the Norwegian and global sample

<table>
<thead>
<tr>
<th></th>
<th>Norwegian sample (N = 44)</th>
<th>Global sample averages (N =450)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>r</td>
<td>r²</td>
</tr>
<tr>
<td>Forceful-enabling versatility</td>
<td>.72**</td>
<td>.52</td>
</tr>
<tr>
<td>Strategic-operational versatility</td>
<td>.65**</td>
<td>.42</td>
</tr>
<tr>
<td>Overall versatility</td>
<td>.71**</td>
<td>.50</td>
</tr>
</tbody>
</table>

*p ≤ .05, **p ≤ .01

A significant relationship between perceived effectiveness and the different versatility scales in the Norwegian sample was found, with r-values ranging from .65 to .72 (p ≤ .01). This entails an explained variance of 42 % to 52 %. Thus, forceful-enabled versatility and strategic-operational versatility explained similar amounts of the variance in perceived effectiveness in the sample. It is interesting how similar the results were to the global sample, with identical values for strategic-operational versatility and overall versatility.
4.4.3 Which form of versatility is the stronger predictor?

The relative contribution of each dimension was also deemed worth exploring. By running a multiple regression analysis, we can examine the predictive power of each of the variables separately and explore the predictive ability of the model as a whole. Before this is possible, the correlation between the independent variables (in this case; forceful-enabling versatility and strategic-operational versatility) must be investigated, to ensure that the results are not affected by multicollinearity. Multicollinearity occurs when some of the independent variables are strongly correlated. High levels of collinearity will increase the standard errors of \( b \) coefficients, limit the size of R and make it difficult to assess the individual importance of a predictor (Field, 2013).

According to Field (2013), very high correlations of above .80 or .90 can be used as a rule of thumb to identify multicollinearity. Forceful-enabling versatility and strategic-operational versatility were found to correlate .88 (\( p \leq .01 \)) in the Norwegian sample, as seen in table 11.

Table 11

| Correlation, variance inflation factor and tolerance for the relationship between forceful-enabling versatility and strategic-operational versatility in the Norwegian sample. |
|---------------------------------|----------------|----------------|
|                                | \( r \) | VIF | Tolerance |
| Forceful-enabling versatility and strategic-operational versatility | .88** | 4.55 | .22 |

*\( p \leq .05 \)  **\( p \leq .01 \)

With such a high correlation, there exists little unique variance, indicating that it would be difficult to discern which variable explains the most variation. The collinearity diagnosis statistics gave a variance inflation factor of 4.55 and tolerance of .22. This means that, according to the collinearity statistics, approximately 22 % of the variance explained by the two factors in the dependent variable was not shared by both variables in this data set. Thus, there was an issue with collinearity in this sample, which affects the interpretation of the results.

Although the predictors are collinear, a multiple regression analysis can still indicate how independent variables \textit{together} predict perceived leader effectiveness and was therefore performed. It is important to underline that the analysis will not give valid results about unique contribution, or concerning which predictors are redundant. A standard multiple regression analysis on leader level, with perceived leader effectiveness as the dependent variable and
forceful-enabling versatility and strategic-operational versatility as independent variables, was computed. The results are presented in table 12.

Table 12
Summary of regression analysis with perceived leader effectiveness as dependent variable in the Norwegian and global sample

<table>
<thead>
<tr>
<th>Type of versatility</th>
<th>Norwegian sample (N = 44)</th>
<th>Average of global samples (N = 450)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Standardized Coefficient Beta</td>
<td>Standardized Coefficient Beta</td>
</tr>
<tr>
<td>Forceful-enabling versatility</td>
<td>.69**</td>
<td>.27*</td>
</tr>
<tr>
<td>Strategic-operational versatility</td>
<td>.04</td>
<td>.51*</td>
</tr>
<tr>
<td>R</td>
<td>.72</td>
<td>.71</td>
</tr>
<tr>
<td>R Square</td>
<td>.52</td>
<td>.50</td>
</tr>
</tbody>
</table>

*p ≤ .05  **p ≤ .01

According to the comparable standardized beta coefficients, there was a big difference in the specific contribution from the two versatility scores. Forceful-enabling versatility had an estimated beta coefficient at .69 (p ≤ .01), while it was estimated to be a non-significant .04 for strategic-operational versatility. Due to the high collinearity the results should be interpreted with caution. The relevant measure is therefore how much variance the independent variables are able to explain together. When combined, forceful-enabling versatility and strategic-operational versatility explained 52.2% of the variance in perceived effectiveness. In other words, versatility accounted for half of what separated the most effective from the least effective managers in this sample.

The proportion of variance explained was approximately the same as was found in the global sample (~50%). In this sample, both types of versatility were found to have a unique contribution to the prediction of perceived effectiveness (Kaiser, 2006). Furthermore strategic-operational versatility was found to be the stronger predictor in the global sample, with an average beta-weight (β = .51) nearly twice as large as for forceful-enabling versatility (β = .27).

4.4.4 The relationship between versatility and effectiveness indicators
As described earlier, Kaiser and Overfield (2010) measure leadership effectiveness by three indicators: the perceived effectiveness of the individual leader as rated by peers, team vitality
rated by subordinates and team productivity as rated by superiors. Table 13 shows the results from three multiple regression analyses from Kaiser and Overfield (2010), using coworkers versatility ratings of 484 leaders to predict the aforementioned effectiveness indicators in the global sample. The analyses were performed on leader level data.

Table 13  
*Summary of regression analysis predicting leadership effectiveness indicators in the global sample.*

<table>
<thead>
<tr>
<th>Effectiveness indicators in the global sample (N = 484)</th>
<th>Perceived leader effectiveness rated by peers</th>
<th>Team vitality rated by subordinates</th>
<th>Team productivity rated by superiors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Forceful-enabling versatility</td>
<td>.30**</td>
<td>.32**</td>
<td>-.02</td>
</tr>
<tr>
<td>Strategic-operational versatility</td>
<td>.38**</td>
<td>.16*</td>
<td>.38**</td>
</tr>
<tr>
<td>Model R</td>
<td>.64**</td>
<td>46.**</td>
<td>.36**</td>
</tr>
</tbody>
</table>

*p ≤ .05 **p ≤ .01

The results show that forceful-enabling versatility was the stronger predictor of subordinate attitudes, while strategic-operational versatility was the stronger predictor of productivity. Finally, both types of versatility affected peer-rated effectiveness.

Due to the challenges of multicollinearity in the Norwegian sample, Pearson's r was used instead of regression analysis in the exploration of the relationship between versatility and the three effectiveness indicators, as assessed by their respective subgroups. The correlations in the Norwegian sample, explored on leader level, are presented in table 14.
Correlations between versatility and leadership effectiveness indicators rated by different subgroups in the Norwegian sample

<table>
<thead>
<tr>
<th>Effectiveness indicators in the Norwegian sample (N = 44)</th>
<th>Perceived leader effectiveness rated by peers</th>
<th>Team vitality rated by subordinates</th>
<th>Team productivity rated by superiors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Forceful-enabling versatility</td>
<td>.64**</td>
<td>.48**</td>
<td>.01</td>
</tr>
<tr>
<td>Strategic-operational versatility</td>
<td>.49**</td>
<td>.41**</td>
<td>-.06</td>
</tr>
<tr>
<td>Overall versatility</td>
<td>.59**</td>
<td>.45**</td>
<td>-.02</td>
</tr>
</tbody>
</table>

*p ≤ .05 **p ≤ .01

There were significant correlations between all three versatility scores and subordinates’ ratings of team vitality, with r-values ranging from .41 to .48 (p ≤ .01). The relationship between versatility and leader effectiveness rated by peers followed the same trend, with even stronger correlation coefficients (.49 - .64). Forceful-enabling versatility had a somewhat stronger correlation with both team vitality and perceived effectiveness compared to strategic-operational versatility, but the differences were small.

For team productivity rated by superiors, there were no significant correlations with the versatility measures. It is important to note that the superior rater group was much smaller than the other two groups, as only one superior - but several peers and subordinates - evaluated every manager. When investigating the correlation between versatility and effectiveness rated by all coworkers, significant correlations (p ≤ .01) between versatility and all the effectiveness indicators were found, ranging from .35 to .72 (see table 15).
Table 15
Correlations between versatility and leadership effectiveness indicators rated by all coworkers in the Norwegian sample

<table>
<thead>
<tr>
<th>Effectiveness indicators in the Norwegian sample (N = 44)</th>
<th>Perceived effectiveness rated by all coworkers</th>
<th>Team vitality rated by all coworkers</th>
<th>Team productivity rated by all coworkers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Forceful-enabling versatility</td>
<td>.72**</td>
<td>.61**</td>
<td>.49 **</td>
</tr>
<tr>
<td>Strategic-operational versatility</td>
<td>.65**</td>
<td>.49**</td>
<td>.35 **</td>
</tr>
<tr>
<td>Overall versatility</td>
<td>.71**</td>
<td>.56**</td>
<td>.43 **</td>
</tr>
</tbody>
</table>

*p ≤ .05 **p ≤ .01

It is important to note that including all ratings will inevitably make common rater effects particularly prominent (Podsakoff, MacKenzie, Lee & Podsakoff, 2003), and so the results should be interpreted with caution.

4.5 Differences in the frequency of leadership behaviors

To explore possible cultural differences relevant for the LVI tool, the degree to which the different leadership behavior dimensions were over- and underdone was explored. In the LVI, scores that fall outside of +/- 3 standard error of the mean of the leadership dimensions are interpreted as over- or underdone (Kaiser, personal communication, January 28th, 2019). The frequencies of overdone and underdone behavior for the four dimensions, analyzed on leader level, are presented in table 16.
There was a higher degree of both overdoing and underdoing enabling and operational behaviors in the Norwegian sample. Overdoing forceful and strategic behaviors, on the other hand, was more common in the global sample. There were also differences in the rank order of the behavior dimensions, as in which dimensions were most under- or overdone.

Table 16
Ranked overview of overdoing and underdoing the four behavior dimensions in the Norwegian and global sample

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Overdone (%)</th>
<th>Underdone (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Norwegian sample (N = 44)</td>
<td>Global sample (N = 18,216)</td>
</tr>
<tr>
<td>Forceful</td>
<td>18 %</td>
<td>24 %</td>
</tr>
<tr>
<td>Enabling</td>
<td>5 %</td>
<td>4 %</td>
</tr>
<tr>
<td>Strategic</td>
<td>2 %</td>
<td>3 %</td>
</tr>
<tr>
<td>Operational</td>
<td>23 %</td>
<td>6 %</td>
</tr>
</tbody>
</table>
5 Discussion

Our intent with this thesis was to investigate the psychometric qualities of the Norwegian translation of the LVI and how it compares to the American version. We also wished to explore the nature of the versatility construct in a Norwegian sample. Overall, the results indicate a satisfactory level of reliability for three out of four dimensions, and for 9 out of 12 subdimensions. However, the subscale e3 supports under the enabling dimension, the subscales o2 efficiency and o3 order under the operational dimension, as well as the operational dimension scale itself, displayed weaker scale reliability. The estimates of both inter-rater reliability and inter-rater agreement showed good results, except for inter-rater agreement for subordinate ratings of the operational dimension.

The validity measures were mixed in their support of the model. Analyzes showed negative correlations between the opposite dimensions. Indications towards curvilinear relationships between the four dimensions and perceived leader effectiveness were also found, although only the relationship with the strategic and forceful dimensions were significant. Explorations of the relationship between the versatility measures and the effectiveness indicators (perceived leader effectiveness, team productivity and team vitality) showed significant correlations when rated by all coworkers.

The theoretical model with four separate dimensions was not replicated in the principal component analysis. The PCA without predetermined components extracted two components for the strategic and operational scales, but only one for the forceful and enabling scales. The three components in this analysis explained 64.1% of the variance in the data. A model with two predetermined components, based on the how and what distinction, displayed a clustering of subscales that partially corresponded with the assumptions of the model. The exception was the operational subscales, clustering closer to origo than expected, and the subscale e1 execution, which grouped together with the forceful subscales. In this analysis, the components explained approximately half of the variance in the data. The confirmatory factor analysis did not support the theoretical model.

Many of the analyses gave results similar to those found for the global samples. The distributions and means of both the versatility scores and scores on effectiveness indicators were similar. All reliability estimates showed comparable values, except for the operational dimension values which were lower. This difference is also similar to trends in earlier studies on the American LVI. In addition, the negative relationship between the opposing dimensions, as
well as the curvilinear relationship between the four dimensions and perceived leader effectiveness were similar to results from the global sample.

An interesting difference between the samples was the predictive validity of the how and what-domains. In the global sample, strategic-operational versatility most strongly predicted perceived leader effectiveness. In our data, forceful-enabling versatility was the strongest predictor, but due to multicollinearity these results are difficult to interpret. The Norwegian data also resulted in a different number of dimensions than what has been found for the global data.

Furthermore, the Norwegian sample had somewhat higher versatility scores than the global sample. The Norwegian leaders also had higher rates of overdoing and underdoing for most behavior dimensions, with the exception of overdoing forceful and strategic behaviors. In the following section, we will discuss the results in the light of both theoretical and methodological explanations and limitations.

5.1 Theoretical Implications

5.1.1 The operational leadership dimension

Throughout the different analyses in this thesis, the operational dimension has been the weak link in the model. The analyses revealed low alpha values for this dimension, markedly lower than for the other dimensions. Furthermore, the operational dimension scale had low inter-item correlations, where some of the correlations were zero or even had negative r-values. Alpha if item deleted also increased for items 8, 11 and 12 under the operational scale. These are strong indicators that the items are not measuring the same underlying construct (Briggs & Cheek, 1986), further supported by the finding of multidimensionality in two out of the three subscales in the operational scale.

The American LVI, however, was found to have alpha values above .70 for the operational dimension and the associated subscales. In addition, the exploratory factor analysis from studies on the American LVI found support for the operational dimension as a distinctive factor in the four-factor model. This could mean that the challenges with the operational dimension in the Norwegian LVI stem from methodological limitations or translation problems in the present study. However, it is worth noting that this dimension has posed the greatest challenge in the American LVI as well, where the operational scale required several rounds of corrections before it reached the psychometric quality it has now (Kaiser et al., 2010). Untangling the causes for
the divergences in the operational dimension in the Norwegian LVI tool will be an interesting line for further research.

5.1.2 Are strategic and operational opposites, or simply complementary?

Our analysis showed that the correlation between the dimensions in the two overarching domains were less clear for the strategic-operational dimension pair than for the forceful-enabling dimension pair. This mirrors the trend in the global sample, where the negative correlation between the strategic and operational dimension is only -.21.

There are also reasons to question the degree to which these dimensions represent opposites when reviewing the correlations between them at item and subscale-level. While the forceful-enabling item pairs for the most part have a negative correlation at -.30 or above, few of the strategic-operational pairs display the same, with one item pair even correlating positively. The subscales display similar patterns, with markedly lower correlations for the strategic and operational subscales. These findings challenge the idea that the strategic and operational dimensions are, in fact, opposites.

That the forceful-enabling dimension pair are opposite but complementary behavior dimensions, where overdoing one type of behavior necessarily excludes the possibility of the other, appears in many ways intuitive. In a situation where a decision needs to be made, it could prove challenging to be both decisive and participative. Therefore, if one type of behavior is overdone, there will naturally be less room for the other to be performed. The dichotomy is less clear for the strategic-operational dimension pair; a behavior could benefit both short-term results and at the same time position the company for future growth.

Another reason for this difference could be that the mechanisms that make leaders neglect one dimension in favor for another is not as strong for strategic-operational. A possible cause described earlier is the human bias of grouping behaviors or attributes into “me” or “not-me” (Kaiser & Overfield, 2010). This mechanism could be stronger in the case of how one relates to other people, than for how one organizes a business. Although no clear conclusions can be drawn based on this limited sample, the results pose an intriguing topic for further exploration.

5.1.3 Factor analyses and the underlying theoretical model

The first principal component analysis resulted in three components, with the forceful and enabling dimensions grouping into one common component. This goes against the assumption of four dimensions specified by the LVI-model. When exploring an alternative interpretation of
the theoretical model, with two predetermined components representing the *how* and *what* distinction, forceful-enabling became the poles of the *how*-component, while the strategic-operational represented the poles on the *what*-component. The majority of the subscales also clustered together as expected, although the operational subscales once again deviated from the expected pattern. This could indicate a weakness in the strategic-operational dichotomy and thus the alternative interpretation of the *what* behavior category as one, single dimension. In addition to the way the operational subscales failed to load sufficiently on the *what*-component, this two-dimensional model also explained less variance and so was not necessarily a better fit than a three-component model for the Norwegian LVI.

The fact that the forceful and enabling dimensions were grouped together as a single dimension raises the question of whether the model is best conceptualized as four-dimensional. On the other hand, studies on large, representative samples have found distinctive patterns in effectiveness scores for too much forceful/too much enabling and too little forceful/too little enabling, a finding typically expected of two dimensions rather than a single, bipolar one (Kaiser, 2019). At the same time, such differences in the theoretical and observed can also be found for a two-dimensional model, as perfectly negatively correlated poles cannot be achieved when error is added to the equation.

The confirmatory factor analysis did not find support for the suggested four dimension-model in the data. The chi-squares were significant, but this is common and is not necessarily a clear rejection of the model (Kenny, 2015). More serious is the fact that the model was below cut-off for the TLI and above for the RMSEA-values for both the forceful-enabling and the strategic-operational pair. Although the results are not a support of the model, they are not a ubiquitous rejection either, as the results could be a product of the sample size or the reliability of the observed variables. Nevertheless, the findings suggest that the problem with the internal structure might not lie in the strategic-operational dimension pair alone, but also affects the forceful-enabling scale. An important caveat to any interpretations is that the analysis was run on rater level data, which means that an analysis on leader level data, e.g. on the relationship between the different leader behaviors rather than rater mindset, could result in a better fit of the model.
As can be seen on the results from both the exploratory and confirmatory factor analysis, there are still several unanswered questions and challenges to whether the theoretical structure of the LVI suggested by Kaplan and Kaiser actually manifests for the Norwegian LVI. These questions concerning the theoretical structure of the model seem worthy of further exploration, to better understand the nature of these opposing, but complementary constructs, especially when seen together with the findings suggesting that strategic-operational are not, in fact, opposites.

5.1.4 **Effectiveness indicators**
Analyses exploring the relationships between the four dimensions and perceived leader effectiveness revealed tendencies towards curvilinear relationships, where the most effective leaders were rated as being closest to “the right amount” (+/-50) for the behavior dimension in question. However, the analyses showed that there were leaders rated as both more and less effective along the whole scale. One should therefore be careful with drawing any conclusions from the results, especially considering the small sample and the restriction of range.

In the global sample, regression analysis revealed that the forceful-enabling pair drives subordinate attitudes, strategic-operational versatility drives team results, and that both versatility measures influence the peer-perceived effectiveness. This can be interpreted as support for the how and what distinction. It also supports the notion that the two types of flexibility have unique functions. Due to multicollinearity, it was not possible to identify the relative contribution of the versatility variables in our sample.

5.1.5 **Is there an underlying trait behind the two forms of versatility?**
There was a high correlation (r = .88) between strategic-operational versatility and forceful-enabling versatility. This means that - in the present sample - leaders who a rated as versatile on the how of leadership typically are perceived as versatile in their use of behaviors related to the what of leadership as well, and vice versa. These results beg the question of whether there is some universal trait behind the behaviors. It would not be surprising if some common factor lies behind leader versatility; it follows the trend of many other human traits, where people who perform well in one area are more likely to perform well in others, for example the g factor of intelligence (Gignac, 2015). Although previous research has found that flexibility as a trait does not fully cover the variation explained by the LVI-model in effectiveness, it could be that some other underlying construct predicts flexibility in behavior.
5.1.6 Differences in degree of overdone and underdone behaviors

As described in the results, although the distributions of the global and Norwegian sample are quite similar, the Norwegian sample has an even higher frequency of leaders who have an overall versatility score above 90%. At the same time, there were a higher percentage of the leaders that were scored as overdoing and underdoing all leadership behaviors except of overdoing forceful and strategic behaviors.

The most overdone leadership dimension globally is forceful behavior, at 24% (Kaiser Leadership Solutions, 2016). In the Norwegian sample, the operational dimension is the most overdone (25%). Previous analyses of top overdone behaviors by region have also found that the forceful behaviors are less overdone in Scandinavia (Kaiser, 2016) compared to the US and other countries. Forceful is the second most underdone dimension in the Norwegian sample, while it is the second least underdone in the global sample. Furthermore, while the Norwegian sample has a higher degree of underdoing overall, the difference between the global and Norwegian sample in underdoing enabling behaviors is quite small (48 vs 44%). In total, there seems to be a smaller bias towards forceful behavior and a higher degree of overdoing operational behaviors in the Norwegian sample, compared to the global sample. There also seems to be a relatively smaller bias towards underdoing enabling behaviors. The fact that Norway is a part of the Scandinavian work culture (Grennes, 2003) could contribute to these differences. In addition, the greatest difference between Norwegian and American culture on Hofstede's cultural dimensions is on the masculinity-femininity scale, which seems more relevant for the forceful-enabling dimension pair (how) than the strategic-operational pair (what) (Hofstede Insights, 2019). Therefore, it seems more in line with previous findings that the differences in the global and Norwegian sample to a greater degree manifest in this dimension pair.

Although the trends in the sample seem to harmonize with previous findings, it is important to remember that another potential reason for such scores is simply the small sample size, as small samples have a higher risk of achieving extreme values (Bornstein, Jager & Putnick, 2013). Therefore, no clear conclusions can be drawn before a larger sample is available.

Although there were some interesting differences, the samples also showed clear similarities in the trends of the distribution. Underdoing was much more common than overdoing in both samples. Furthermore, in both samples, forceful was clearly the most overdone behavior dimension, strategic the least overdone dimension and operational the least underdone
dimension. This indicates that there are some challenges to leader versatility that have cross-cultural relevance. Why these aspects of leadership are so challenging is an interesting question to explore in other studies.

It could be argued that the American and Norwegian cultures are so similar that it does not pose a true examination of cross-cultural validity. Nevertheless, as described earlier, there exist important differences that are especially relevant for organizational culture. In future studies, it may be relevant to compare the findings with trends in data from other nations as well, to ensure an even better understanding of how versatility manifests in different cultures.

5.1.7 Effects of femininity on correlations between versatility and effectiveness indicators

Although the regression analyses were unstable due to multicollinearity, it is still interesting to see that it is the how-factor that becomes the most powerful predictor of the effectiveness indicators in the Norwegian sample - the opposite of what has been found in the global sample. Is it possible that the highly feminine Norwegian work culture results in a workplace where it is necessary to master the art of relations in order to be an effective leader? When seen in light of the gender ratio as well, it could be that Norwegian organizational culture is more dependent on classically “feminine” skills than other cultures. The ratio of men and women in the Norwegian sample was more balanced than in the global sample, with 51% versus 28% female leaders.

A counter-argument to a possible influence is that previous studies on gender and the LVI have not found any systematic differences between the genders in LVI scores nor in leadership style (Kaiser & Wallace, 2016). Neither did they find indications that female leaders were evaluated differently (Kaiser & Wallace, 2016). Nevertheless, being leaders in one of the worlds’ most gender equal cultures, in addition to being surrounded by many other female leaders, might affect both the female leaders and their male colleagues’ leadership style in unknown ways. Hopefully, further studies will shed more light on this topic.

5.2 Limitations and future research

There are a number of methodological limitations to this study. Many of them can be remediated through further research, but it is important to review the limits to conclusions based on the data.

5.2.1 Sample

A major limitation to any assumptions or interpretations in this study is the small sample size. This is especially true for analyzes on leader level. A sample of 44 managers is merely a fraction
of the number of leaders used to validate the American version of the LVI. Such a small sample increases the risk that the population represented is a very special one, with low levels of generalizability. Small samples are also more vulnerable to errors, outliers and extreme scores, since they can be drawn from very specific populations. The results are also less likely to be replicated, as the population the sample is drawn from easily changes (Button et al., 2013). Thus, there is a problem of both instability and generalizability with small samples. Data from more leaders and from a more diverse pool of businesses is needed before the conclusions can be generalized.

5.2.2 Public sector

Another aspect of the sample that makes it different to the global sample is the large proportion of leaders who work in the public sector. Trends in the data could therefore be caused by effects specific for the type of organizations included in the study. The public sector is said to differ markedly from the private sector on a number of factors, such as job security, competitiveness and strategic focus. This could create challenges as to the usefulness of the LVI-model, which was developed in a culture where the public sector is much less prominent than in Norway. An example is that strategic growth is a meaningless measure for most leaders in the Norwegian public sector. Following our research, these differences have been taken into account and Henning Bang and colleagues has developed an alternative LVI version for the public sector. Further studies on this new, adapted version could perhaps address some of the issues described here.

5.2.3 Translation

There will always be issues related to the translation and adaptation of assessment tools. The goal for translations of psychometric instruments is to maintain the conceptual meaning, rather than the linguistic equivalence (Harkness, Pennell & Schoua-Glusberg, 2004). This can lead to challenges for several different reasons, including cultural differences in how one expresses oneself, and linguistic differences such as the lack of well-known Norwegian equivalents to American words, e.g. "effectiveness" and "efficiency". Future studies may benefit from comparing both the English and Norwegian versions in the same population, in order to further evaluate the quality of the translation.

5.2.4 Restriction of range

The descriptive data revealed a severe restriction of range in the Norwegian sample. In, addition, several of the analyses performed used averages from rater groups, as well as averages across
items in subscales. Averages will naturally utilize only parts of the range in the scale, thus possibly increasing the effect of range restriction. Restrictions of range can reduce the size of correlations and coefficient alpha (Sackett, Laczo & Arvey, 2006). The advantage of aggregate scores is that it increases the reliability compared to using single-item scores. In a sample as small as ours, this is especially useful. Therefore, it was seen as the best way to treat the data.

Restriction of range in the versatility scores could, in addition to the score being a compound score, be due to an artifact of the LVI-tool, e.g. a ceiling effect, where the tool does not sufficiently measure variation between leaders performing on a very high level. The leaders in this sample may be from an especially high-performing population. Businesses and leaders willing to engage in time-consuming leadership assessments as part of a research-project might be more interested in development and reflection around leadership behaviors than average. Therefore, the businesses included in this thesis could have especially strong leaders, which would make it harder to discern differences in performance at the very end of the available spectrum.

5.2.5 Common method bias

Another issue with the data is common method bias, where some of the variation could be a function of common measurement and/or the source used to gather the data (Podsakoff et al., 2003). Common rater effect is a potential source of common method bias, which in this thesis would come into play for example when ratings of versatility and effectiveness done by the same rater are used as both the independent and dependent variable in regression analyses. As humans tend to evaluate things and people holistically, this may inflate the relationship between the variables. A remedy to this issue would be to compare the versatility scores of half of the raters with the effectiveness-scores of the other half of raters. Due to the small sample size, this remedy was not performed for the present sample.

Social desirability has also been cited as a source of common method bias that can curb potential relationships and lead to spurious correlations (Ganster, Hennessey & Lothans, 1983). There is a risk that employees do not answer honestly because of potential unpleasantness, or due to fear of negative consequences for giving negative feedback. Although the raters are anonymous and the responses are grouped together in the LVI tool, there is still a possibility that such an effect is present and so should be heeded.
In multi-rater assessments, a single rater often evaluates several different leaders and will thereby influence several lines of data. Normally, one rater will not be involved in the assessment of too many leaders and so will not unduly affect the analyses of the data, but in a sample as small as ours, such a bias could have significant effects on the overall results.

5.2.6 Sampling
A convenience sample is vulnerable to selection bias and sampling errors, which constrains the generalizability and power in the study (Bornstein et al., 2013). Convenience sampling is suitable for pilot studies and hypothesis generation, as well as for data collection over short durations of time (Bornstein et al., 2013). For these reasons, it can be seen as appropriate for the scope of the present study, but it entails need for caution when interpreting the results. The fact that the data was collected over a short time period also makes it vulnerable to the influence of specific events in the environment, e.g. a recent conflict that will affect the responses in the questionnaire. At the same time, it is important to remember that most of psychological research utilizes some form of convenience sample.

5.2.7 Recommendations for future research
The LVI-survey is based on coworker-reports and self-assessment, and does not contain any objective measures of leader effectiveness or other outcome variables. What has been found is a distinct relationship between leader versatility, as measured by coworkers, and coworkers’ perception of the leader’s effectiveness. Furthermore, although there has been established a relationship between versatility and the effectiveness indicators, we can not draw causal conclusions about the relationships between variables based on correlations alone (Shadish, Cook & Campbell, 2002).

In addition, a number of mediating and moderating factors such as work experience, leadership level, age and demographic variables, and confounding variables such as personality, could influence the observed results. Further research could therefore benefit from controlling for such variables.

As this was a comparative study and the studies on the American LVI utilized Cronbach’s alpha to estimate scale reliability, this was seen as the right choice for our thesis as well. Nevertheless, as the dimension scales are composed of several subscales and the assumption of item equivalence is not necessarily met, Cronbach’s alpha might not be the best estimate of scale reliability for the four dimensions.
Future research should take into account the aforementioned limitations and adapt their research design accordingly. Performing longitudinal studies and exploring the effect of improved versatility on perceived leader effectiveness - achieved through leader development programs - could also add new insight and alleviate a number of the biases present in this study.

5.3 Practical Implications
As has been elaborated upon above, caution must be used when interpreting the results of this study. Nevertheless, the results show that versatility as measured by the Norwegian LVI correlates with perceived leader effectiveness, team productivity and team vitality, which are relevant parameters for the success of businesses. This indicates that versatility is an important part of being a successful manager in Norway as well, and that more is not necessarily better. Rather, Norwegian leaders could do better pursuing the right amount of the different leadership behaviors.

As the tool is innovative in its use of a non-linear scale to measure behavior and in its focus on versatility, it could enrich our understanding of leadership if the reliability of the tool is improved. Since the operational scale was not found to give stable scores, great caution should be used when interpreting the results from this scale and the subscales in particular. Due to the weaknesses described above, important decision regarding leaders and their further development should not be based on the current version of the Norwegian LVI alone.
6 Conclusion

Overall, the results indicate a satisfactory level of reliability for the forceful-enabling dimension pair. The strategic-operational dimension pair displayed considerable weaknesses in both reliability and internal structure, as the relationship suggested by the theory did not manifest as expected. The operational dimension is the main weakness of the LVI-model in the Norwegian LVI. The factor analyses did not find support for the four-dimensional model as described in the theory. The findings should be investigated further to increase the psychometric quality of the tool and better the understanding of the nature of the versatility construct in Norwegian populations. It is important to note that, although some severe weaknesses were found, this does not automatically discredit the use of the Norwegian LVI, as there are significant methodological limitations to the study.

This is the first cross-cultural comparison and documentation of the applicability of the LVI in a sample where the majority is not American leaders. That the results - as well as the weaknesses in the model - were as similar to the global results as they were is striking. Taken together, this could offer a cautious indication that the LVI could have cross-cultural validity in practical application.
7 References


Hofstede Insights. Compare countries - USA and Norway. Retrieved from


Innovasjon Norge (n.d.). Business Culture. Retrieved from


Kaplan, R. E., & Kaiser, R. B. (2010). Toward a positive psychology for leaders.


8 Appendix

Appendix 1

*Inter-item correlation and alpha if item deleted for the operational scale*

<table>
<thead>
<tr>
<th>Item nr.</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>Alpha if item deleted</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.67</td>
<td></td>
<td></td>
<td>.67</td>
</tr>
<tr>
<td>2</td>
<td>.37</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.61</td>
<td></td>
<td></td>
<td>.61</td>
</tr>
<tr>
<td>3</td>
<td>.46</td>
<td>.46</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.57</td>
<td></td>
<td></td>
<td>.57</td>
</tr>
<tr>
<td>4</td>
<td>.26</td>
<td>.33</td>
<td>.38</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.66</td>
<td></td>
<td></td>
<td>.66</td>
</tr>
<tr>
<td>5</td>
<td>.06</td>
<td>-.08</td>
<td>.12</td>
<td>-.14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.33</td>
<td></td>
<td></td>
<td>.33</td>
</tr>
<tr>
<td>6</td>
<td>.02</td>
<td>-.02</td>
<td>.06</td>
<td>-.08</td>
<td>.52</td>
<td></td>
<td></td>
<td></td>
<td>.36</td>
<td></td>
<td></td>
<td>.36</td>
</tr>
<tr>
<td>7</td>
<td>-.04</td>
<td>.10</td>
<td>.12</td>
<td>-.07</td>
<td>.27</td>
<td>.33</td>
<td></td>
<td></td>
<td>.40</td>
<td></td>
<td></td>
<td>.40</td>
</tr>
<tr>
<td>8</td>
<td>.17</td>
<td>.14</td>
<td>.25</td>
<td>.20</td>
<td>.04</td>
<td>-.07</td>
<td>.14</td>
<td></td>
<td>.65</td>
<td></td>
<td></td>
<td>.65</td>
</tr>
<tr>
<td>9</td>
<td>.04</td>
<td>.11</td>
<td>.11</td>
<td>.06</td>
<td>.13</td>
<td>.07</td>
<td>.09</td>
<td>-.02</td>
<td></td>
<td>.24</td>
<td></td>
<td>.24</td>
</tr>
<tr>
<td>10</td>
<td>.12</td>
<td>-.16</td>
<td>.02</td>
<td>-.22</td>
<td>.33</td>
<td>.41</td>
<td>.09</td>
<td>-.16</td>
<td>.15</td>
<td></td>
<td>.50</td>
<td>.50</td>
</tr>
<tr>
<td>11</td>
<td>.17</td>
<td>.06</td>
<td>.15</td>
<td>.48</td>
<td>-.09</td>
<td>-.01</td>
<td>.17</td>
<td>.21</td>
<td>.13</td>
<td>-.17</td>
<td></td>
<td>.54</td>
</tr>
<tr>
<td>12</td>
<td>-.00</td>
<td>-.07</td>
<td>.16</td>
<td>-.01</td>
<td>.26</td>
<td>.17</td>
<td>.18</td>
<td>.06</td>
<td>.45</td>
<td>.23</td>
<td>.23</td>
<td>.11</td>
</tr>
</tbody>
</table>
## Appendix 2

*Numbers of factors extracted for the dimensions and subscales in the Norwegian sample*

<table>
<thead>
<tr>
<th>LVI Dimensions</th>
<th>Subscales</th>
<th>Norwegian sample</th>
<th>Number of factors extracted (eigenvalue above 1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Forceful</td>
<td>F1 Takes charge</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>F2 Declares</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>F3 Pushes</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Enabling</td>
<td>E1 Empowers</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>E2 Listens</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>E3 Supports</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Strategic</td>
<td>S1 Direction</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>S2 Growth</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>S3 Innovation</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Operational</td>
<td>O1 Execution</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>O2 Efficiency</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>O3 Order</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>
## Appendix 3

*Correlations between opposites, item-level*

<table>
<thead>
<tr>
<th>Item</th>
<th>Mean</th>
<th>SD</th>
<th>Corrected item-total correlation</th>
<th>Correlation between opposites</th>
<th>Item</th>
<th>Mean</th>
<th>SD</th>
<th>Corrected item-total correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1f. In control</td>
<td>-0.24</td>
<td>0.92</td>
<td>0.51</td>
<td>-0.31</td>
<td>1e. Empowers people</td>
<td>-0.17</td>
<td>0.84</td>
<td>0.73</td>
</tr>
<tr>
<td>2f. Assumes authority</td>
<td>-0.22</td>
<td>0.92</td>
<td>0.48</td>
<td>-0.30</td>
<td>2e. Gives people room</td>
<td>-0.08</td>
<td>0.64</td>
<td>0.63</td>
</tr>
<tr>
<td>3f. Gives direction</td>
<td>-0.22</td>
<td>0.89</td>
<td>0.57</td>
<td>-0.37</td>
<td>3e. Hands-off</td>
<td>0.04</td>
<td>0.72</td>
<td>0.67</td>
</tr>
<tr>
<td>4f. Steps in</td>
<td>-0.03</td>
<td>0.81</td>
<td>0.30</td>
<td>-0.30</td>
<td>4e. Trusts people</td>
<td>-0.01</td>
<td>0.76</td>
<td>0.58</td>
</tr>
<tr>
<td>5f. Decisive</td>
<td>-0.03</td>
<td>1.06</td>
<td>0.68</td>
<td>-0.14</td>
<td>5e. Participative</td>
<td>-0.36</td>
<td>0.88</td>
<td>0.64</td>
</tr>
<tr>
<td>6f. Takes stands</td>
<td>-0.18</td>
<td>1.00</td>
<td>0.74</td>
<td>-0.20</td>
<td>6e. Relies on input</td>
<td>-0.26</td>
<td>0.76</td>
<td>0.75</td>
</tr>
<tr>
<td>7f. Speaks up</td>
<td>-0.24</td>
<td>0.94</td>
<td>0.61</td>
<td>-0.29</td>
<td>7e. Open to influence</td>
<td>-0.24</td>
<td>0.79</td>
<td>0.63</td>
</tr>
<tr>
<td>8f. Doesn't back down easily</td>
<td>0.16</td>
<td>0.91</td>
<td>0.51</td>
<td>-0.35</td>
<td>8e. Receptive to push-back</td>
<td>-0.33</td>
<td>0.77</td>
<td>0.65</td>
</tr>
<tr>
<td>9f. Pushes people hard</td>
<td>-0.16</td>
<td>0.81</td>
<td>0.56</td>
<td>0.02</td>
<td>9e. Shows appreciation</td>
<td>-0.29</td>
<td>0.74</td>
<td>0.30</td>
</tr>
<tr>
<td>10f. Expects a lot</td>
<td>-0.06</td>
<td>0.66</td>
<td>0.39</td>
<td>-0.02</td>
<td>10e. Treats people well</td>
<td>-0.06</td>
<td>0.49</td>
<td>0.55</td>
</tr>
<tr>
<td>11f. Direct when dissatisfied</td>
<td>-0.40</td>
<td>0.83</td>
<td>0.57</td>
<td>-0.32</td>
<td>11e. Sensitive to people's feelings</td>
<td>-0.08</td>
<td>0.76</td>
<td>0.53</td>
</tr>
<tr>
<td>12f. Holds people accountable</td>
<td>-0.49</td>
<td>0.84</td>
<td>0.61</td>
<td>-0.37</td>
<td>12e. Gives the benefit of the doubt</td>
<td>0.06</td>
<td>0.77</td>
<td>0.27</td>
</tr>
<tr>
<td>Item</td>
<td>Mean</td>
<td>SD</td>
<td>Corrected item-total correlation</td>
<td>Item</td>
<td>Mean</td>
<td>SD</td>
<td>Corrected item-total correlation</td>
<td></td>
</tr>
<tr>
<td>-------------------------------------------</td>
<td>------</td>
<td>-----</td>
<td>----------------------------------</td>
<td>-----------------------------------</td>
<td>------</td>
<td>-----</td>
<td>----------------------------------</td>
<td></td>
</tr>
<tr>
<td>1s. Future-oriented</td>
<td>-0.36</td>
<td>0.90</td>
<td>0.71</td>
<td>1o. Short-term focus</td>
<td>-0.01</td>
<td>0.71</td>
<td>0.41</td>
<td></td>
</tr>
<tr>
<td>2s. Thinks strategically</td>
<td>-0.33</td>
<td>0.88</td>
<td>0.86</td>
<td>2o. Tactical</td>
<td>-0.08</td>
<td>0.73</td>
<td>0.52</td>
<td></td>
</tr>
<tr>
<td>3s. Big-picture perspective</td>
<td>-0.26</td>
<td>0.76</td>
<td>0.77</td>
<td>3o. Attention to detail</td>
<td>-0.09</td>
<td>0.83</td>
<td>0.57</td>
<td></td>
</tr>
<tr>
<td>4s. Anticipates change</td>
<td>-0.32</td>
<td>0.77</td>
<td>0.77</td>
<td>4o. Follows up</td>
<td>-0.31</td>
<td>0.67</td>
<td>0.42</td>
<td></td>
</tr>
<tr>
<td>5s. Aggressive about growth</td>
<td>-0.28</td>
<td>0.92</td>
<td>0.61</td>
<td>5o. Conservative about risk</td>
<td>0.15</td>
<td>0.81</td>
<td>0.42</td>
<td></td>
</tr>
<tr>
<td>6s. Launches many changes</td>
<td>-0.52</td>
<td>0.90</td>
<td>0.63</td>
<td>6o. Practical about change</td>
<td>-0.01</td>
<td>0.76</td>
<td>0.42</td>
<td></td>
</tr>
<tr>
<td>7s. Bold moves</td>
<td>-0.46</td>
<td>0.92</td>
<td>0.70</td>
<td>7o. Incremental change</td>
<td>-0.06</td>
<td>0.71</td>
<td>0.37</td>
<td></td>
</tr>
<tr>
<td>8s. Entrepreneurial</td>
<td>-0.40</td>
<td>0.87</td>
<td>0.74</td>
<td>8o. Efficient</td>
<td>-0.15</td>
<td>0.75</td>
<td>0.06</td>
<td></td>
</tr>
<tr>
<td>9s. Questions the status quo</td>
<td>-0.35</td>
<td>0.81</td>
<td>0.51</td>
<td>9o. Goes by the book</td>
<td>0.03</td>
<td>0.64</td>
<td>0.37</td>
<td></td>
</tr>
<tr>
<td>10s. Embraces change</td>
<td>-0.19</td>
<td>0.60</td>
<td>0.65</td>
<td>10o. Stays with tried and true</td>
<td>0.11</td>
<td>0.67</td>
<td>0.10</td>
<td></td>
</tr>
<tr>
<td>11s. Open to new ideas</td>
<td>-0.16</td>
<td>0.60</td>
<td>0.68</td>
<td>11o. Organized</td>
<td>-0.30</td>
<td>0.70</td>
<td>0.09</td>
<td></td>
</tr>
<tr>
<td>12s. Encourages innovation</td>
<td>-0.29</td>
<td>0.67</td>
<td>0.67</td>
<td>12o. Process-oriented</td>
<td>-0.02</td>
<td>0.57</td>
<td>0.51</td>
<td></td>
</tr>
</tbody>
</table>
## Appendix 4

*Correlations between opposites, subscale-level*

<table>
<thead>
<tr>
<th>Scale</th>
<th>Mean</th>
<th>SD</th>
<th>Mean inter-item correlation</th>
<th>α</th>
<th>Correlation between opposites</th>
<th>Scale</th>
<th>Mean</th>
<th>SD</th>
<th>Mean inter-item correlation</th>
<th>α</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1</td>
<td>-0.18</td>
<td>0.63</td>
<td>0.34</td>
<td>0.68</td>
<td>-0.49</td>
<td>E1</td>
<td>-0.06</td>
<td>0.59</td>
<td>0.54</td>
<td>0.82</td>
</tr>
<tr>
<td>F2</td>
<td>-0.07</td>
<td>0.78</td>
<td>0.52</td>
<td>0.81</td>
<td>-0.35</td>
<td>E2</td>
<td>-0.29</td>
<td>0.65</td>
<td>0.56</td>
<td>0.84</td>
</tr>
<tr>
<td>F3</td>
<td>-0.27</td>
<td>0.59</td>
<td>0.41</td>
<td>0.74</td>
<td>-0.28</td>
<td>E3</td>
<td>-0.10</td>
<td>0.47</td>
<td>0.31</td>
<td>0.61</td>
</tr>
<tr>
<td>S1</td>
<td>-0.31</td>
<td>0.72</td>
<td>0.69</td>
<td>0.90</td>
<td>0.09</td>
<td>O1</td>
<td>-0.12</td>
<td>0.53</td>
<td>0.36</td>
<td>0.70</td>
</tr>
<tr>
<td>S2</td>
<td>-0.42</td>
<td>0.73</td>
<td>0.56</td>
<td>0.84</td>
<td>-0.32</td>
<td>O2</td>
<td>-0.02</td>
<td>0.51</td>
<td>0.21</td>
<td>0.52</td>
</tr>
<tr>
<td>S3</td>
<td>-0.25</td>
<td>0.54</td>
<td>0.52</td>
<td>0.81</td>
<td>-0.19</td>
<td>O3</td>
<td>-0.05</td>
<td>0.40</td>
<td>0.18</td>
<td>0.44</td>
</tr>
</tbody>
</table>
### Appendix 5

*Rotated component matrix, PCA with two predetermined components*

<table>
<thead>
<tr>
<th></th>
<th>Component 1</th>
<th>Component 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1 Takes charge</td>
<td>.80</td>
<td>.25</td>
</tr>
<tr>
<td>F2 Declares</td>
<td>.73</td>
<td>.17</td>
</tr>
<tr>
<td>F3 Pushes</td>
<td>.65</td>
<td>.32</td>
</tr>
<tr>
<td>E1 Empowers</td>
<td>-.71</td>
<td>.23</td>
</tr>
<tr>
<td>E2 Listens</td>
<td>-.47</td>
<td>.57</td>
</tr>
<tr>
<td>E3 Supports</td>
<td>-.55</td>
<td>.24</td>
</tr>
<tr>
<td>S1 Direction</td>
<td>.26</td>
<td>.80</td>
</tr>
<tr>
<td>S2 Growth</td>
<td>.30</td>
<td>.86</td>
</tr>
<tr>
<td>S3 Innovation</td>
<td>.04</td>
<td>.87</td>
</tr>
<tr>
<td>O1 Execution</td>
<td>.52</td>
<td>.04</td>
</tr>
<tr>
<td>O2 Efficiency</td>
<td>.01</td>
<td>-.40</td>
</tr>
<tr>
<td>O3 Order</td>
<td>.05</td>
<td>-.27</td>
</tr>
</tbody>
</table>
Appendix 6

Visual representation of model fit for the LVI model in the global sample

Items/Behaviors | Subdimension | Dimension
--- | --- | ---
Takes charge | Takes Charge | .83
Assumes authority | .77
Steps in | .56
Sure of self | .73
Declares self | .73
Doesn't back down | .74
Pushes people hard | Pushes | .73
Sets stretch goals | .72
Holds people accountable | .70
Empowers; able to let go | Empowers | .63
Trusts people | .83
Receptive to others' ideas | Listens | .94
Open to influence | .76
Makes it easy to push back | .74
Provides support | Supports | .68
Shows appreciation | .75
Compassionate | .77