Evaluation of Measurement Invariance of the Student Self-Reported Learning Outcome Measure:

Validation of the Norwegian Student Survey

Lucy Wairimu Gitiria

Centre for Educational Measurement at the University of Oslo (CEMO)

Master of Science in Evaluation, Measurement, and Assessment

MAE 4090: 30 Credits Master Thesis

Professor Dr Rolf Vegar Olsen

Spring, 2023.

Popular Abstract

The Norwegian Agency for Quality Assurance in Education's student self-reported learning outcomes measure allows 5th and 2nd-year students in Norwegian higher education institutions to report on their satisfaction with learning outcomes. The aggregate scores from the measure are used to offer relevant comparative information concerning these indicators, to institutions offering higher education, applicants to higher education, the government, students, and other educational stakeholders. The current use and interpretation of these scores imply unidimensionality of the factor structure and invariance of the structure across study groups. However, for such claims to hold, an empirical justification is required, for the use and interpretations of information from the scale to be valid. This study, therefore, contributes to gathering this evidence by evaluating the scale's factor structure, its generalizability (reliability) and measurement invariance across four selected study program types. The implied unidimensionality was not supported. Evidence for comparability of the factor structure across study programs is supported partially. Three items in the scale were found to be scalar non-invariant. These findings to some extent can be used to justify the revision of the scale to better inform the interpretations and use of the scores by the users of the information.

Acknowledgments

I duly acknowledge and thank my supervisor Professor Dr Rolf Vegar Olsen for initiating and giving me a chance to write on this topic. I am grateful for his insightful comments and suggestions throughout the writing of this thesis.

I also recognise and thank the Norwegian Agency for Quality Assurance in Education for allowing us to use their scale for this study and the Norwegian Agency for Shared Services in Education and Research for providing access to the data that enabled the writing of this thesis.

Special thanks to the CEMO staff and the AME 2021 cohort for your feedback and continued support.

To my family – thank you for the inspiration, emotional and financial support throughout my study period.

Abstract

The Norwegian Agency for Quality Assurance in Education's student self-reported learning outcomes measure assesses the learners' satisfaction with their learning outcomes using ten generic items. The aggregate scores are used to offer relevant comparative information concerning these indicators, to institutions offering higher education, applicants to higher education, the government, students, and other educational stakeholders. To draw valid inferences from this construct, especially concerning the comparability of higher education study programs, an inspection of the psychometric properties, including the validity of the measure, is necessary. Based on the current use of the scores a unidimensional model is expected to fit the data. Therefore, confirmatory factor analysis is conducted using robust maximum likelihood to determine the extent to which data from the 2018 cycle of the student survey revealed the learning outcomes measures' structure. The plausibility of comparability claims is evaluated using measurement invariance tests done across four selected study program types. A study sample composed of respondents from the nursing program (2194), business and administration program (2952), teacher education (1032) and engineering program (1310), who answered all items on the learning outcomes measure was used. The implied unidimensional model was not supported. The data supported a modified single-factor model. Multigroup confirmatory factor analysis test results supported configural and metric invariance, indicating equivalence of the latent concept and structure across the four study groups. Full scalar invariance was not achieved, however, after releasing intercept equality constraints of three items, partial scalar invariance was achieved. Accurate and valid measurement of learning outcomes is crucial for people that depend on the scale to make important decisions. These findings can be used to initiate a revision of the scale to ensure confident comparisons across groups.

Keywords: Measurement invariance, Self-reported learning outcomes, Validation, Norwegian student survey, Higher education.

Evaluation of Measurement Invariance of the Student Self-Reported Learning Outcome Measure: Validation of the Norwegian Student Survey

The need to improve competitiveness, transparency, recognition, and mobility in higher education, resulted in a focus on learning outcomes (LOs) as a means of assessing knowledge, skills, and competences in education (Adam, 2006). Learning outcomes, defined as what a learner has achieved and can be demonstrated at the end of the learning activity (Prøitz 2010) – represent a practical and methodological approach to achieving these goals. In Europe, the motivation to shift toward learning outcomes in higher education was reinforced by initiatives such as Bologna Process, the Tuning Process, and the European Qualification Framework for Lifelong Learning (EQF). National Qualification Frameworks (NQF) alongside the Standards and Guidelines for Quality Assurance in the European Higher Education Area (ESG) are being used to offer coordinative opportunities, emphasize the use of learning outcomes, and ensure efficient ways of measuring LOs at the country level. The qualification frameworks and the ESG consider learning outcomes as a point of quality reference in higher education. It is presumed that a close link exists between how an educational institution describes and facilitates its learning outcomes, and the quality of a program (Adam, 2006; Hansen et al., 2013).

As a response to the shift towards learning outcomes and its importance as a quality indicator in higher education, the Norwegian Agency for Quality Assurance in Education (NOKUT) introduced the learning outcome measure in the annual student survey – Studiebarometeret. Its formulation was guided by the Norwegian National Qualification Framework for Lifelong Learning (NQF) (P. Bakken, personal communication, September 29,2022). Specifically, the student self-reported learning outcomes (SSRLO) measure, assesses the learners' satisfaction with their learning outcomes using ten generic items. The aggregate scores are used to offer relevant comparative information concerning these indicators, to institutions offering higher education, applicants to

higher education, the government, students, and other educational stakeholders (Norwegian Agency for Quality Assurance in Education [NOKUT], 2019).

To draw inferences from this construct, especially concerning the comparability of higher education study programs, an inspection of the psychometric properties, including validity of the SSRLO measure, is necessary. Based on the theory of validity, researchers establish validity arguments by considering among others, the evidence on the factor structure, and the comparability of the factor structure across groups (Standards for Educational and Psychological Tests [AERA] et al., 2014; Kane, 2013; Marsh, 1994). However, we have not come across a study investigating the factor structure of the Norwegian student survey's SSRLO measure. In addition, to the best of our knowledge, the comparability of the SSRLO measure across different subgroups (e.g., study program) has not been established – even though sufficient comparability evidence ought to be established before valid group comparisons are made.

This study, therefore, contributes to the literature by investigating the factor structure of the SSRLO measure. Furthermore, we evaluate the extent to which the factor structure is invariant across selected study programs. So that, any differences across these groups can be evaluated. The focus and contribution of this study, therefore, is in the discussion of SSRLO's construct validity - specifically measurement invariance of the measure.

This paper is organised as follows: we present the conceptual and theoretical framework in the next section. Then the methodology, comprising of data and sample, a brief description of the measure, and data analysis - specifically, handling missing data, factor structure testing, reliability, and invariance testing. The last section presents the results, a discussion of the results, limitations and implications of the study and a conclusion.

Conceptual and Theoretical Framework

The concept of Learning Outcomes in Higher Education

Several definitions of LOs exist in literature – differing depending on the author's perspectives of learning and their intent of measuring the LOs. Based on theories rooted in behaviourist perspectives such as objectives movement and curriculum planning, some authors theorize LOs as predefined statements of expected or anticipated outcome to be demonstrated by the learner's performance. These established definitions emphasize LOs in the context of curriculum development and their realization/assessment (Prøitz, 2010). Yet others based on open-ended perspectives such as constructivists, critic the aspect of predefining LOs arguing that prespecified LOs cannot cover all learning. As such the alternative definitions offered differ in their formulation, depending on the author's area of interest. However, they all generally present LOs as what a learner ends up with- acquired through interaction with meaningful material in and beyond higher education (HE) institutions (Caspersen et al., 2011; Prøitz, 2010).

Despite this diversity, two clear overall approaches to defining learning outcomes emerge in most literature: a) the curriculum approach – with a focus on program objectives i.e., what needs to be taught and the teaching strategies therein and b) the individual learning approach – with a focus on the measurement of what the student should have attained after learning. This attainments can further be related to institutional effectiveness (Adam, 2006; Allan, 1996; Caspersen et al., 2011; Prøitz, 2010). With the introduction of national qualification frameworks and the consequential shift of focus to student centred learning, emphasis is now on the actual learners' achievements rather than what is being or is to be taught (Adam, 2006). As such, one of the commonly accepted definition of LOs especially in Europe is - "statements of what a learner is expected to know, understand and/or be able to do at the end of a period of learning" (The European Credit Transfer System [ECTS], Users' Guide, 2005 p.47). Here, learning outcomes are seen as what the learner should know and can do because of the learning process (Adam, 2008).

The ECTS Users' Guide definition of LOs is the most relevant to the aims of this study given it emphasises the measurement of what the student has learnt and can be demonstrated at the end of the learning activity. NOKUT measures students' satisfaction with their LOs at second and fifth year of study, i.e., what they have achieved and can demonstrate from their fields of study - up until the time of assessment.

Although a common ground seems to be found with the establishment of the qualification frameworks, existing literature shows that discrepancies still exist concerning approaches of assessing LOs in HE – i.e., how, and core dimensions (what) of assessment.

Approaches of Assessing Learning Outcomes in Higher Education

There are different ways of measuring LOs mainly determined by the construct being measured, the goal of measuring and the level of measurement (institutional, program, or course). So, the measure may either be objective or self- reported (Pike, 1996). Conventionally, objective measures such as grades and standardized tests have been seen as sufficient measures of learning. NOKUT for instance assesses the achievement of learning outcomes using standardized topic specific national tests for specific study programs like nursing[1]. However, with increased need for comparability and transparency across European HE institutions and the consequent focus on LOs, (Stensaker & Sweetman, 2014), standardized measures, more so grades have been found impractical – especially regarding comparisons across disciplines and institutions. Self-reported surveys have been argued to offer an alternative to grades.

In higher education, emphasis is particularly being placed on measuring LOs that can be attributed to quality of higher education programs (Douglass et al., 2012). As such, aggregated ratings from self-reported learning outcomes (SRLOs) are being used to compare educational practices and make major decisions across different fields (Kuh, 2005). However, due to the subjective nature of many self- reported measures of outcomes, their validity has often been

[1] See https://www.nokut.no/utdanningskvalitet/nasjonal-deleksamen/

questioned (Gonyea, 2005). Thus, some researchers and policy makers tend to trust direct measures more compared to self-reports.

Studies shows that SRLO are credible and valid measures of learning, with SRLO items varying across academic groups as theoretically predicted (Pike, 2011). Contrary, other studies have questioned the construct validity of SRLO questions, citing cognitive inability of students to accurately report their learning outcomes. In addition, some findings indicate low correlations between objective measures and self-reported measures of LOs (Bowman, 2011; Caspersen, Smeby, et al., 2017; Porter, 2013). Other studies have found differences in some SRLO dimensions across programs and professions (Caspersen et al., 2014). Explanations given in literature for the differences observed in the relationship between direct and SRLO measures could be attributed to the nature of the measures – one measures the actual skill whereas the other measures the perceptions of the students concerning their skills (Pike, 1996). Other explanations include introduction of self-biases (Allen & Van Der Velden, 2005; Gonyea, 2005; Pike, 1993; Tourangeau et al., 2000); unclear concepts or scales without proper anchors (Ouimet et al., 2004), and the difference in scope of measurement (Pike, 1996). Despite this differences, SRLO use should not be completely ignored. If well designed, they offer alternatives to measuring and understanding LOs in higher education (Douglass et al., 2012).

In literature, there are diverse perspectives on core dimensions of LOs that should be assessed. In a comparative review of assessment of LOs in HE, Nusche (2008) presents a categorization of LOs indicators based on evidence from eighteen assessment instruments. The typology presents LOs as a multidimensional aspect highlighting cognitive and non-cognitive outcomes. According to Nusche, *cognitive learning outcomes* refers to recognition of knowledge and intellectual abilities that a learner has developed. This may span from specific to more broader thinking and problem-solving processes and are mostly based on the classical and later improved Bloom's (1956, 1997) taxonomy. Cognitive LOs assessment can emphasise either knowledge acquirement or skills acquirement. *Knowledge acquirement* indicators can either be discipline

specific i.e., the knowledge acquired is from a specific field (e.g., Mathematics) or general knowledge acquired independent of one's field. Knowledge acquirement needs learners to simply remember ideas, theory, or given materials and phenomena (pp. 8-9). Dias and Soares (2017) also suggested a similar classification that they refer to as hard skills – that can either be applied/theoretical specific knowledge or applied/theoretical generic knowledge.

Nusche (2008) defined *cognitive skills acquirement* as the ability of a student to put in use the knowledge acquired to solve real life problems and complete tasks. It mainly includes acquirement of transferable skills that transcend study programs. In HE, assessments that aim to compare LOs across disciplines use generic transferable skills. The indicators include e.g., verbal communication, quantitative reasoning, problem solving, critical and innovative thinking, processing of information and evaluation of ideas depending on the student's level (pp. 9-10). This is similar to what Kraiger et al. (1993) referred to as skill based learning and what Dias and Soares (2017) refer to as generic soft and transferable skills.

Beyond the acquisition of cognitive knowledge and skills, HE institutions also emphasise acquisition of *non-cognitive skills*. Nusche (2008) defines *non-cognitive skills* as changes that occur in an individuals' beliefs or/and values due to interactions with core learning or through extracurricular activities organised by higher education institutions. The most assessed non-cognitive learning includes psychosocial development (e.g., interpersonal relations), attitudes, and values (e.g., social responsibility and respect for diversity). Yet, the choice of specific non-cognitive indicators to assess may still be challenging e.g., due to the varying cultures and beliefs.

Given the distinct nature of higher education and the broad definition of LOs, it has been argued that generic SRLO scales tends to be more flexible to apply in different settings in higher education(Caspersen, Smeby, et al., 2017; Nusche, 2008). Since they do not focus on specific disciplines or professions, same generic SRLO scale can be used to compare students across different study programs and institutions. Methodologically, generic SRLO cover a broader array of subject matter and can be electronically dispensed. From a practical perspective, they are cost effective and

easy to use (Gonyea, 2005), thus, popular among researchers in different fields. Yet there are still some doubts whether generic LOs can be attributed to only the university learning experiences. Generic skills are rarely an explicit part of course study programs outcomes, therefore the role that HE institutions play in their acquirement is not clear. Ewell (1991; as cited in Nusche, 2008) suggests that generic skills acquirement may be because of social maturation and not necessarily learning experiences in higher education. Therefore, there is a risk of assessing aspects beyond higher education influences in learning.

Scales with generic learning outcomes indicators are in use in various parts of the world. Their focus and items used vary depending on the purpose of measurement. In their review of seven prominent educational and workforce frameworks, Markle et al. (2013) identified seven common micro categories of generic transferable skills indicators that can be utilized in an assessment framework – i.e., creativity, critical thinking, teamwork, effective communication, information technology, citizenship and life skills. Other studies like Tremblay et al. (2012) in the Assessment of Higher Education Learning Outcomes (AHELO) feasibility study; Burrus et al. (2013) from a work-force oriented view; and Oswald et al. (2004) in a review of several higher education institutions' students learning outcomes and mission statements – all identified similar domains of generic learning outcomes.

The SRLO Measure across groups

There are exacerbated concerns about instrumentation when measurements are provided by self-reports. More so when the attributes being measured are not observable like intentions, beliefs, and attitudes. This is because this kind of instruments are made through Likert scales, which are prone to response biases like social desirability and acquiescent responding (Gonyea, 2005). Given the variability of self-reported responses in each population, overlooking their possible effects on the score may compromise validity and reliability of the assessment.

Most SRLO measures are attitudinal - representing the individual subjective beliefs about their learning outcomes. As such, individual or contextual differences may arise in the measure

making interpretation of intergroup findings unclear. It becomes challenging to establish whether the differences are due to true attitudinal differences or due to varied psychometric responses to the items in the measure. This is of particular concern in higher education research when systematic differences in the students' perceptions about their learning outcomes are due to institutional/ discipline differences or academic differences (Caspersen et al., 2014). It is argued that the differences are to some extent based on the social organisation of knowledge networks within the disciplines (Caspersen, Frølich, et al., 2017). The knowledge base of a discipline and the beliefs emphasising the socialisation of students determines the learning strategies that are emphasised. Emphasis in hard pure disciplines like natural sciences are completely different from those in hard applied disciplines (e.g., science-based professionals) like Nursing (Muller, 2009). This may reflect in how the student responds to the SRLO items. In Norwegian higher education for instance, Engineering and Nursing studies are based on two distinct educational policies, and varying academic traditions. The academic orientation differs but both are considered hard disciplines with a focus on practical application of knowledge (Caspersen et al., 2014).

The theory of person-environment fit by Holland (1997) can also partly be used to explain the possible differences that may arise among students of different academic backgrounds. The theory posits that individuals can be organised into one or more of Holland personality types i.e., Realistic, Investigative, Artistic, Social, Enterprising and Conventional. It is also argued that there are corresponding six model environments exhibiting dominant societal setting - under which individuals can be categorised. Each environment has individuals with personality types that allow them to engage in, utilize their competences and skills, and express their values and attitudes. The theory by Holland, emphasises that different academic majors socialize and reinforce students' abilities differently and these contextual differences may show in the interpretation of the SRLO items (Caspersen et al., 2014). It is therefore important to investigate whether potential differences exist across selected groups, when using self-reports, for valid comparisons to be made.

Validation Process

For inferences from a construct to be valid, especially concerning comparability of scores across groups, psychometric properties of the measure, specifically the factorial structure of the measure and measurement invariance must be inspected. Validity of a measure depends on how the scores will be interpreted and used (AERA et al., 2014). Therefore, any validation study must consider the purpose and use of the scores from the measure. The validation process assesses whether the claims being made concerning the proposed interpretations and uses of scale score are warranted given the empirical evidence (Kane, 2013). In addition, validation may call for revisions of the interpretations and uses of the score.

The student surveys in the Norwegian context like other prominent student surveys with generic SRLO items e.g., the National Survey of Student Engagement (NSSE), provide learning institutions and other stakeholders with information regarding learning in the different institutions. In addition, they offer comparative information on generally how different campuses compare in terms of attaining the LOs. Furthermore, they act as a rich source of data that researchers use to understand different phenomena in the educational sector (Pike et al., 2012). Therefore, every single use should be validated.

This thesis is consequently a validation of the Norwegian student surveys' SRLO measure - with a focus on the claims of intended use of scores i.e., to offer comparative information concerning students' satisfaction with their LO. The information from the measure is presented as a composite score that is used to compare study programs on the different aspects presented by the indicators. The use of composite scores (sum scores) assumes that variation on every indicator is caused by a single general factor (unidimensionality) (McDonald, 1999). The comparisons made using the sum scores across study programs presumes equivalence of the factor structure (Vandenberg & Lance, 2000). The plausibility of these claims is thus evaluated by examining the factor structure, the generalizability (reliability) and the invariance of the measure across the selected groups.

Assessing Factorial Structure

Few studies have examined the structural component of self-reported data. However, general guidelines for evaluation of factor structure are available in literature. It is advisable to have a strong theory. Without a strong theory, factor analysis of the measures is not appropriate (Brown, 2015). However, some authors in support of self-reported data argue that with a clear definition of the construct, it can be ascertained whether the self-reported data represents the construct being measured (Gonyea, 2005; Pike, 2011).

Before conducting a validation study, its therefore key to identify the construct that the scale is supposed to measure (Pike, 1992). The construct of interest is usually defined by the domains allegedly measured by the scale. Assessing the factor structure of a measure therefore regards establishing how many latent factors are measured and whether it corresponds to dimensions of the concept (Brown, 2015). In other words, the researcher seeks to determine whether the inter-relationship of the indicators supports the intended scores used to draw inferences. For example, if a test is meant to report a composite score- then one-dimension is expected. Identifying and or verifying the underlying dimensions and the pattern of factor loadings of scores generated from a measure - requires a clear understanding and specification of the theoretical or conceptual framework of the construct.

Reliability

As a component of validity evidence related to internal structure, reliability evaluates any inconsistency of scores originating from differences among individuals on the construct (Kane, 2013). It is expected that scores that measure the same factor should exhibit high reliability values. Any systematic variation in the scores that measure the same construct indicates internal inconsistency (McDonald, 1999). Cronbach's alpha coefficient is commonly used to indicate the internal consistency of a measure- and many studies provide this coefficient. However, given its stringent assumptions (e.g., one-dimensionality), some researchers prefer the MacDonald omega-

which relaxes such strict assumptions (Revelle & Zinbarg, 2009). When the evidence of factor structure and reliability is confirmed, invariance of the measure across groups can be assessed.

Measurement Invariance

Measurement invariance (MI) is describing whether the structure of a measurement instrument is equivalent across diverse groups. Equivalence in this case is linked to the measurement level where scores obtained in different groups are comparable. The classification of (Van de Vijver & Leung, 1997) presents three forms of equivalence. Construct or structural equivalence refers to measuring the same construct in every group. This confirms whether the model holds for all the selected groups. Measurement unit equivalence involves ensuring same measurement units across the groups, but the origin unit may be different. This confirms whether the scale intervals are the same, by constraining the factor loadings/ slopes to equality. If established, it implies that the unstandardized regression coefficients and covariances can be compared between groups. Finally, full score equivalence is ensuring same unit and origin of the scores. This has implication on the comparability of the latent means across the groups (Vandenberg & Lance, 2000).

Comparability of scores across groups is mostly affected by bias. Bias is systematic measurement errors that creates other explanations of intergroup differences (Van de Vijver & Tanzer, 2004). It could be differences in the underlying meaning of the construct across groups, bias due to methodological procedures used or variations of a tool at item level (Van de Vijver & Tanzer, 2004). Bias threatens the validity of intergroup comparability and therefore all sources of bias and error should be minimized. Equivalence of the measurement level has a significant function in comparing groups and when MI is not supported across different groups, it is impossible to interpret findings that show differences in this groups.

When testing the invariance concerning the psychometric properties of measurement instrument, the common invariance tests conducted include – configural, metric, and scalar invariance tests. Residual invariance which relates to the premise that the sum of measurement

error and specific variance of the items is the same across all groups (Vandenberg & Lance 2000), is not commonly tested especially for studies that aim to compare latent means (Putnick & Bornstein, 2016). For this reason, this study focused only on configural, metric, and scalar invariance tests.

Configural invariance means that data for each group exhibit the same number of factors with the same set of indicators related to each factor (Meredith, 1993). It provides evidence that respondents from different groups perceive the structure of the construct in the same way. If the concepts are abstract, or respondents from different groups attach different meaning to the construct, configural invariance will not be established. It is key that configural invariance is established for other MI tests to be meaningful (Vandenberg & Lance 2000).

Metric invariance means that the factor loadings are equal across the groups. This implies that the groups calibrate the measure in the same way – thus values on the observable scale have same meaning across the groups. According to Steinmetz et al. (2009) factor loadings can be interpreted as validity coefficients since they signify the strength of the causal effect of the latent variable on the items. Therefore, having factor loading equal in both groups, depict that the measure has the same meaning and structure across groups - a prerequisite for meaningful cross-group comparison (Vandenberg & Lance 2000).

Scalar invariance means the equivalence of the item factor loadings and intercepts (values of every item equivalent to zero value of the construct). It implies that mean differences in the construct being measured reflects all mean differences in the shared variance of the indicators (Cheung & Rensvold 2002; Putnick & Bornstein, 2016). If the item intercepts significantly differ in one group, it is an indication that there are some systematic response biases. Hence non scalar invariance, and thus latent mean comparisons will be ambiguous. This is because the differences in latent means observed between the groups is affected by scale and origin of the latent variable differences (Vandenberg & Lance 2000). If invariance of the configural, metric and scalar models is achieved, latent means can be compared.

Full invariance as described above can be challenging to achieve in practice. Therefore, Byrne et al.'s (1989) idea of partial invariance (invariance of a subset of the parameters) has been adopted in some research (Vandenberg & Lance 2000).

The key assumption in all MI research is that the structure of the measure used is equivalent across the groups being compared. The validity of this assumption is vital for conclusions to be made concerning group differences (Vandenberg & Lance, 2000). In addition, if this assumption is not confirmed, it is not certain that the construct being measured is the same across the groups (Steinmetz et al., 2009). Thus, for valid comparison of means or the structural relations between groups, the measurement structures underlying the indicator must be equivalent.

The Present Study

The Norwegian Student Survey's self-reported learning outcomes (SSRLO) measure has been in use since the 2014 cycle. However, no empirical study has been conducted to establish the psychometric properties of the measure or its generalizability across groups. The present study's purpose is to contribute to formulating a validity argument for the current interpretation and use of SSRLO measure (sum scores - compared across study program types). Psychometric properties of SSRLO are thus examined through confirmatory factor analysis (CFA). CFA is chosen because it can assess both the factor structure and invariance of all measurement parameters of the model across several groups.

First, CFA is used to examine the SSRLO structure. Based on NOKUT's conceptualization of LOs and the current use and interpretation of the scores from this scale, a single factor model is implied. Therefore, we posit that a single factor model represents the data. Second, given that the SRLO measure may have considerable differences across study programs, we investigate the MI, and the significance of mean differences is tested between the groups. So far, nothing is known about the MI of the measure, thus, establishing MI to a sufficient degree may enable confident mean comparisons between the groups, to better inform the potential users of the information generated by this tool. Invariance is tested across four study group types that are likely to indicate large

differences. We limit the invariance testing to the aspects of the measurement i.e., configural, metric and scalar invariance.

This study contributes to the literature on validity evidence based on internal structure and enhance understanding of measurement invariance for the NOKUT SSRLO scale across study programs.

Research Questions

The research will be guided by the following questions:

1. To what extent does the data from respondents of the 2018 cycle of the student survey reflect the SRLO factor structure?
    a. Which factor structure fits the data better based on the intended use of the scale scores?
2. To what extent does the measurement model of the SSRLO scale show invariance across the four selected study program types?

Method

Data and Sample

The Norwegian student survey (Studiebarometeret) 2018 data is used in this analysis. This is a national survey organised by NOKUT. In 2018, more than 30,000 students from about 1800 study programs participated in the sixth cycle of the survey. The participants were selected from all 40 institutions of higher learning with bachelor's and master's programs. The data was collected cross sectionally between October and November 2018 using self-reported questionnaires that are both in English and Norwegian. A total of 31,256 students completed and returned the questionnaires. Of this about 37.37% were male and 62.63% female.

The Norwegian student survey data used was obtained from the Norwegian Agency for Shared Services in Education and Research[2] (SIKT). NOKUT has classified the students' reported study programs into 39 study program types. For ease of analysis this variable was used to select

[2] Data can be obtained on request from https://www.sikt.no

program types that were included in the study. Four groups namely nursing, business and administration program, teacher education, and engineering were selected. The study programs types selected offer a good representation in terms of their varied backgrounds – as either hard/soft, applied/pure, or professional programs (Muller, 2009). In addition, they exhibit diversity in terms of their focus and emphasis on knowledge, skills, and competencies (Caspersen et al., 2014), and are represented by a large number of respondents in the survey. Further, for ease of interpreting MI results, we avoided having too many groups, thus the four representative groups were preferred. Overall missing data due to non-response on one or more of the ten learning outcome variables was approximately 17% (i.e., 19% for business and administration group; 14% for engineering group; 15% for the nursing group and 14% for the teacher education group. Therefore, information on LOs was available for most of the students. To ensure the same subset of cases is used in the analysis, listwise deletion of the missing values was used (Peugh & Enders, 2004). Consequently, the study sample composed of respondents from the nursing program (2194), business and administration program (2952), teacher education (1032), and engineering program (1310). The study was approved by SIKT (see Appendix I).

The Measure

The NOKUT learning outcomes scale is an attitudinal measure within the national student survey, where students gauge their level of satisfaction with learning outcomes. According to NOKUT, the formulation of the items was guided by the national qualification framework. Students reported on 10 items (see Table 1) and the prompt question was "*How satisfied are you with your learning outcomes so far, concerning*". Respondents decide to what degree they were satisfied with their LOs on a 5-point Likert scale ranging from 'Not satisfied' (1) to 'Very satisfied' (5). Specific items in the scale are presented in Table 1 below. Most responses seem to be stable around the value 4.

Table 1

Response Distribution of NOKUT's SSRLO measure (n=7488)

Item	Wording	Mean	Standard Deviation
Item1-	Theoretical knowledge	3.73	0.88
Item2-	Knowledge of scientific work methods and research	3.23	1.02
Item3-	Experience with research and development work	2.99	1.06
Item4-	Discipline- or profession-discipline specific skills	3.43	1.01
Item5-	Critical thinking and reflection	3.83	0.90
Item6-	Cooperative skills	4.02	0.88
Item7-	Oral communication skills	3.73	0.99
Item8-	Written communication skills	3.82	0.87
Item9-	Innovative thinking	3.59	0.97
Item10-	Ability to work independently	4.05	0.89

Note. NOKUT= The Norwegian Agency for Quality Assurance in Education; SSRLO= the student self-reported learning outcomes measure.

Statistical Analysis

Factorial Structure Analysis

Before modelling, multivariate outliers check was done using Mahalanobis distance (Tabachnick et al., 2007) (the 205 outliers identified were not removed, and a sensitivity test was conducted). Multivariate descriptive statistics (multivariate skewness and kurtosis) were conducted using Mardia's (1970) tests. The criteria by Kline (2016) of absolute skewness value of ≤ 3 and kurtosis value of ≤ 10 suggesting less severe non-normal distributions were used. Pairwise scatter plots were used to investigate the linearity assumptions (see Appendix C). All items in the scale were linearly associated.

Confirmatory factor analysis (CFA) was conducted to examine the factor structure of the SSRLO measure. First, based on how the scores from the scale are currently used (sum scores - compared across study programs), unidimensionality of the scale is implied, thus a model with all the items loading on a single factor was specified. The factor loading of the first item was fixed to one.

Based on the framework for assessing LOs in higher education defined by Nusche (2008), the items in the NOKUT scale seemed to capture various sub-dimensions of cognitive LOs. It is therefore

reasonable to say that some items can be classified under *'knowledge acquirement'* and others under '*skills acquirement'*. A model with two factors was fitted to the data with items 1-3 loading on the factor '*knowledge acquirement*' and items 4-10 loading on the factor '*skills acquirement'* (See Table 1). The factor loadings of the first item were fixed to one and the factors were allowed to correlate.

The literature recommends the use of generic transferable skills if the assessment intends to compare LOs across disciplines (Caspersen, Smeby, et al., 2017; Gonyea, 2005; Nusche,2008). Therefore, focusing on this a reduced single-factor model was estimated based on Nusche's framework- items under '*skills acquirements'* (i.e., items 4-10). Factor loading of the first item was fixed to one. The best-fitting model is selected for further MI analysis. Single-group confirmatory factor analysis (SGCFA) was used to examine if the proposed factor structure provides a good fit for the separate groups, to allow MI to be conducted.

Data analysis was conducted in R software (R. Core Team, 2020) using the laavan package (Rosseel, 2012), psych package (Revelle, 2017), and semTools (Jorgensen et al., 2020). Initial normality checks revealed large multivariate kurtosis (mardia's multivariate *kurtosis* =159.28, $p <$ .001). This violates the multivariate normality assumption of the commonly used CFA estimator maximum likelihood (ML). Therefore, to prevent possible standard errors under-estimations, Robust maximum likelihood estimator (MLR) was used for factor analysis since the estimator is suitable for multivariate non-normal data (Rhemtulla et al., 2012).

Given the sensitivity of chi-square tests to sample and model size other different fit indices were considered adequate in the assessment of model fit: Comparative fit index (CFI) and Tucker-Lewis Index TLI ≥.95 (good fit), >.90 (acceptable fit); Root mean square error of approximation (RMSEA) ≤.05 (close fit), between .05 to .08 (reasonable approximate fit) and above .10 (poor fit); Standardised root mean residual (SRMR) close to .08 or below (Brown, 2015; Kline, 2015). Standardized regression loads were used to establish the strength of the factor loadings and values ≥.4, were considered acceptable (Kline, 2014).

Reliability

To establish internal consistency reliability of the responses of the SRLO measure, both the Omega (McDonald, 2013) and Cronbach alpha (Cronbach, 1951) reliability coefficients were calculated. Both were preferred as Omega is deemed better at detecting heterogeneity among items(Revelle & Zinbarg, 2009). For valid use of the total score, the measure is expected to exhibit high reliability. The following criteria was used: Reliability >.9 – excellent, between .8 and .89 – good and .7 – acceptable reliability (McDonald, 1999).

Measurement Invariance Testing

To establish whether the scores from the SSRLO measure are comparable across study programs, and if indeed the observed mean differences between the study groups are meaningful, measurement invariance testing was conducted using multi-group CFA (MGCFA). Specifically, the measurement model for configural invariance, metric invariance and scalar invariance were tested. A configural invariance model – with no constraints placed on the parameters is tested to check the number of factors and pattern of free and fixed parameters in each group. This model forms the basis against which the other consequent models are tested. Attaining configural invariance guarantees the test of metric invariance. For metric invariance, a model with only the factor loadings constrained to be equal, yet the intercepts are allowed to differ across the groups is examined. If metric invariance is achieved, tests of scalar invariance can be conducted. To test the invariance of the intercepts (scalar invariance), a model with the factor loadings and intercepts constrained to be equal across the groups is examined.

Measurement invariance model fit is typically assessed using the chi-square (χ^2) and alternative fit statistics RMSEA, SRMR, CFI, and TLI (Brown 2015; Kline, 2015). The models are evaluated by comparing two nested models. Most researchers recommend the use of 'significant chi-square differences' to evaluate two nested models. However, as noted in the literature, the chi-square is very sensitive to even the slightest deviations from an ideal model, especially in large samples (Chen, 2007; Cheung & Rensvold, 2002). As the sample size increases, the χ^2 has also been

found to increase in power to reject the null hypothesis. Alternatively, a change in CFI (ΔCFI) of ≤ −0.01 in nested models comparing two to three groups is recommended (Chen, 2007). Rutkowski and Svetina, (2014) found that as the number of groups in the MI test increased, ΔCFI decreased and ΔRMSEA increased. For models with more than three groups, and having a large sample size, they proposed less stringent criteria (ΔCFI ≤ - 0.02) for metric invariance. The ΔCFI ≤ - 0.01 remains for scalar invariance. Therefore, this study did not use chi-square differences – instead alternative fit indices suggested by Rutkowski and Svetina (2014) (i.e., ΔRMSEA ≤ 0.03, ΔCFI ≤ - 0.02 and ΔSRMR ≤ 0.03) for metric model and (ΔCFI ≤ - 0.01, ΔRMSEA ≤ -0.015 and ΔSRMR≤ -0.015) for the scalar model was used.

Partial invariance tests as part of the measurement model refers to a less strict test where some of the parameters are allowed to be freely estimated across groups (Putnick & Bornstein, 2016). The partial invariance function in R software was used to conduct partial invariance testing. The parameters of the scalar model were extracted and inspected to identify the items whose intercepts were functioning differently. In this case intercepts of item 4, – 'Discipline and profession-specific skills', was the first identified as the potential source of invariance, then Item 7 – 'Oral communication skills', and finally item 6 – 'Cooperative skills. The group partial function was used to release intercept equality constraints on these items- one at a time – first, item 4 intercept was freely estimated, then items 4&7 intercepts and finally items 4,6 & 7 intercepts freely estimated. All the models were compared against the metric invariance model. The CFI and RMSEA changes were inspected. The suggested criteria for concluding partial invariance – though not empirically supported – is that at least more than half the items on the latent factor must be invariant (Putnick & Bornstein, 2016; Vandenberg & Lance, 2000).

Results

Descriptive Statistics

The response distributions are presented in Table 1. The item means and standard deviations indicate that most respondents are satisfied with their LOs, with a mean ranging from 2.99 to 4.05 (*SD*= 0.88 to 1.02). Item correlations indicate that all variables in the measure were inter-correlated (see correlation matrix in appendix IIIB). Most items exhibited a medium positive association with each other. The highest correlation was between item 2 and item 3 (*r*=.71) whereas the lowest correlation was between item 3 and item 10 (*r*=.28).

Factorial Structure

To determine the factor structure of SSRLO, CFA was conducted, and the results are presented in Table 2. Based on the current interpretations and use of the scores from the scale (sum scores - compared across study programs), we expected that a model with one factor will represent the factor structure. However, a CFA model with all 10 items on the measure exhibited poor fit (TLI= .78; CFI= .83; RMSEA=.122 and SRMR= .071) (see Table 2 – Model A). This indicates that the items as presented in the current scale seem to reflect a multi-dimensional construct. The conceptually proposed two-factor model exhibited a poor fit considering the selected fit criteria (TLI=.88; CFI= .91; RMSEA=0.087, and SRMR= .064) (see Table 2 – model B). Investigation of the modification indices suggested a covariance between item 2- 'Knowledge of scientific work methods and research' and item 3- 'Experience with research and development work'. This may suggest that there is some variation common to the two items not accounted for. This could be due to the overlap of perceptions or similarity of the item's wording/content (use of the word 'research'), thus the two factors in the model did not account for the possibility that respondents had similar answers within these two items –missing an important feature of the data. After a revision of the model as suggested above, the model fit improved (TLI=.91; CFI= .94; RMSEA=0.076, and SRMR= .045) with a factor correlation of (*r*=.79) (see Table 2 – model Bi).

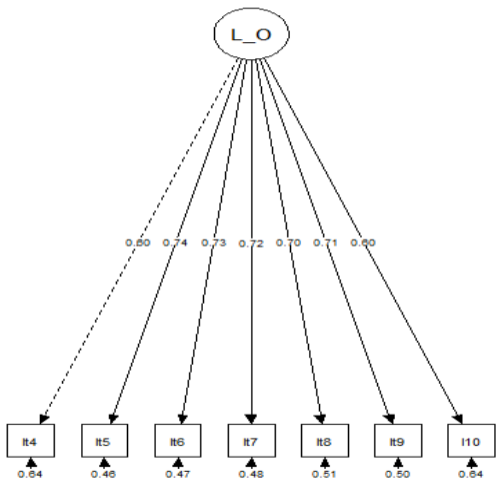
NOKUT intends to compare satisfaction with LOs across study programs in Norwegian HE. As recommended in the literature, we focused on generic transferable skills domains of LOs to enable us to test for invariance later. Items 1 'theoretical knowledge', item 2- 'Knowledge of scientific work methods', and item 3- 'Experience with research and development work' were in the next step excluded from the analysis since they did not meet the criteria (Markle et al., 2013; Nusche, 2008). A unidimensional model – with 7 items was tested. The model fit was within an acceptable range (CFI= .96; TLI= .94; RMSEA=.070 and SRMR= .033) (see Table 2 – Model C). The factor loadings were strong ranging between .60 and .74 (see Table 3 and Figure 1). The measure showed acceptable reliability with an overall Cronbach Alpha (α= .86) and Omega for model C (ω=.86). Model C had a satisfactory fit and was used in subsequent MI analysis.

The final factor structure (Model C) was fit on data for each group to examine if it provides a good fit for the separate groups (business and administration, engineering, teacher education program and nursing program). All the models fit the data well except for the Engineering group – which had fit indices slightly below the set criteria (see Table 2). The standardized factor loadings were all statistically significant and salient (ranging from .56 to .78) (Table 3). Based on the residuals and the modification indices, no significant local strain was seen in the solutions. Therefore, Model C was used further in the MI analysis.

Figure 1

Empirical one factor Model C and the Theoretically implied T*wo Factor Model A*

Model C: *Empirical* Model



Model Bi: Theoretical Implied Model

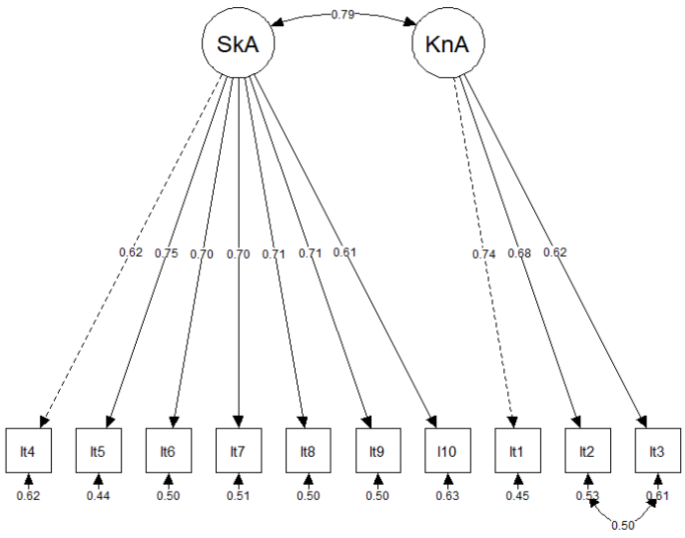


Table 2

Fit Statistics from Confirmatory Factor Analysis Models Tested (N=7488)

Model	**X^2 (df)**	**CFI**	**TLI**	**RMSEA (90% CI)**	**SRMR**
Model A: Single factor model (10 items)	3904.7 (35)	0.828	0.779	0.122 (0.119 0.124)	0.071
Model B: Two-factor model	1951.1 (34)	0.915	0.887	0.087 (0.084 0.090)	0.064
Model Bi: Two-factor model (one error variance)	1461.2 (34)	0.937	0.914	0.076 (0.073 0.079)	0.045
Model C: Single factor model (7 items)	534.3 (14)	0.960	0.940	0.070 (0.066 0.075)	0.033
Single-group Solution (Model C)					
Business & Administration *(n= 2952)*	253.7 (14)	0.954	0.932	0.076 (0.069 0.083)	0.036
Engineering (*n=1310*)	169.7 (14)	0.929	0.893	0.092 (0.082 0.103)	0.046
Teacher Education (*n=1032*)	77.3 (14)	0.963	0.944	0.066 (0.054 0.079)	0.033
Nursing (*n=2194*)	92.0 (14)	0.981	0.972	0.050 (0.043 0.058)	0.023

Table 3

Standardized CFA Factor Loadings for the overall One Factor Model (model C) and Single group CFA (n=7488; α= .86; ω=.86)

Item	Standardized Factor Loadings				
	Overall	Single group CFA of model C			
	Model C	B/Admn	Engineering	Teacher ed	Nursing
Discipline and profession-specific skills	.60	.58	.57	.56	.64
Critical thinking and reflection	.74	.76	.69	.71	.72
Cooperative skills	.73	.72	.69	.71	.77
Oral communication skills	.72	.69	.72	.72	.78
Written communication skills	.70	.72	.72	.72	.69
Innovative thinking	.71	.74	.68	.67	.70
Ability to work independently	.60	.59	.58	.62	.69

Note. Model C= Overall single factor model (7 items); Model C fit on groups (B/Admn=business and administration group data; Engineering group; Teacher Education group, and Nursing group.

Measurement Invariance

To examine the extent that the measurement model of the SSRLO scale shows invariance across the four selected study program types, the CFA model (Model C) was used. Results for measurement invariance models are shown in Table 4. The fit indices for the configural model were acceptable (CFI= .96; TLI= .94; RMSEA=.072, and SRMR= .030) (see MI. configural model in Table 4). This signifies that the model with one factor (7 items) was suitable for each study program type since the structure revealed is similar in the study groups. The alternative fit indices selected for this study indicated support for metric invariance (see MI. metric model in Table 4) i.e., invariant factor loadings across study program types (ΔCFI=-.005, ΔRMSEA -.006). This demonstrates that the compositions of the SSRLO structure are similar for the selected study groups. Compared to the metric model, the scalar invariance model with equal factor loadings and equal intercepts fits the data worse (ΔCFI= -.057, ΔRMSEA =0.024) (see MI. scalar model in Table 4). This demonstrates variability in item intercepts between study programs.

To identify the items that were functioning differently, the model was inspected for poorly fitting parameters using partial invariance testing. The results for partial invariance are presented in Table 5. The change in CFA after releasing intercept equality constraints on item 4 – *'Discipline and profession-specific skills'* (Partial.scalar1 model in Table 5), and items 4 & 7– '*Oral communication skills',* (Partial.scalar2 model in Table 5) was still above the set criteria. Thus, partial scalar invariance was not achieved in these models. Partial scalar invariance was achieved when intercepts of item 4 item 7, and item 6 – '*Cooperative skills*' were allowed to vary freely across the study programs (Partial.scalar3 model in Table 5). Non-invariance in these items suggests some systematic group differences that are not attributed to satisfaction with LOs and the exhibited structure.

Achieving partial scalar invariance can permit the comparison of the groups on the latent mean of LOs (Putnick & Bornstein, 2016). The latent means were accordingly estimated using the model that was demonstrated to be partially invariant – the model where intercepts of three items were freely estimated (Partial.scalar3 in Table 5). The estimated means for each group – were the business and administration group (M=.080), Engineering group (M=.003), teacher education group (M=-.126), and nursing group (M=-.150). The results of the pairwise comparison of the estimated means is presented in Table 6 in Appendix III-G). The results indicate that there is a statistically significant difference between the mean scores of the groups except between teacher education and nursing groups - the test revealed a t statistic of 1.40, with df = 14974 (p-value = 0.1604). The effect size in the compared groups was negligible except between business and administration and nursing groups where the effect size was small with a Cohen d of 0.22.

Table 4

Fit Statistics from the Measurement Invariance Tests on Study Program Variable

Model	df	CFI	RMSEA (90% CI)	SRMR	Model comp.	Δdf	ΔCFI	ΔRMSEA	ΔSRMR	Decision
MI.configural: Configural Invariance	56	0.960	0.072 (0.067 0.076)	0.030	-	-	-	-	-	**Invariant**
MI.metric: Metric Invariance	74	0.956	0.066 (0.062 0.070)	0.038	Model MI1	18	**-0.005**	**-0.006**	0.008	**Invariant**
MI.scalar: Scalar Invariance	84	0.898	0.089 (0.086 0.093)	0.062	Model MI2	10	**-0.057**	**0.024**	0.024	Non invariant

Note. Δ (delta)=change; Model comp.=Model compared; N=7488 (Business & Administration group n=2952; Engineering group=1310: Teacher education group=1032; Nursing group=2194)

Table 5

Fit Statistics for Scalar Partial Invariance Testing

Model	df	CFI	RMSEA (90% CI)	SRMR	Model comp.	Δdf	ΔCFI	ΔRMSEA	ΔSRMR	Decision
MI.metric: Metric Invariance	74	0.960	0.066 (0.062 0.070)	0.038	Model MI1	18	**-0.005**	**-0.006**	0.008	**Invariant**
Partial.scalar1: Scalar partial invariance	89	0.920	0.080 (0.077 0.084)	0.052	Model MI2	15	**-0.035**	**0.014**	0.014	Non invariant
Partial.scalar2: Scalar Partial invariance	86	0.930	0.071 (0.067 0.074)	0.044	Model MI2	12	**-0.015**	**0.005**	0.006	Non invariant
Partial.scalar3: Scalar partial invariance	83	0.945	0.069 (0.065 0.073)	0.043	Model MI2	9	**-0.008**	**0.002**	0.003	**Invariant**

Note. Δ (delta)=change; Model comp.=Model compared; N=7488 (Business & Administration group n=2952; Engineering group=1310: Teacher education group=1032; Nursing group=2194). **Model Partial.scalar1**=intercepts for item 4 freely estimated across groups; **Model Partial.scalar2**=intercepts for item 4 and 7freely estimated across groups; **Model Partial.scalar3**=intercepts for item 4, 7 and 6 freely estimated across groups.

Discussion

The purpose of this study was to validate the current use and interpretations of NOKUT's SSRLO scale scores. The SSRLO provides learning institutions and other stakeholders with information regarding learning in the different HE institutions. NOKUT presents this information as aggregate scores that compare satisfaction with LOs across different study programs in Norwegian higher education. In this case, a unidimensional factor structure and its equivalence across study programs is implied. To generate empirical evidence concerning these claims (Kane, 2013), we investigated the extent to which the data from the 2018 cycle fit the SSRLO structure (Research question 1) and then evaluated measurement invariance across selected study program types (Research question 2).

Based on the current use of the SSRLO we expected that a single-factor model would fit the data. The CFA results show that this was not achieved as a single-factor model with all ten items fit the data poorly. This suggests that the scale as presented now captures more than one dimension. In line with establishing the structure that fits the data well, Nusche's (2008) framework suggests that the indicators span the *knowledge* and *skills* dimensions. The two-factor model estimated fit the data after the inclusion of a covariance parameter between Items 2 and 3 – indicating that the measurement error may not be random between these two indicators. In this case, other scoring methods need to be considered since a typical sum score cannot be applied (McDonald, 1999). In addition, the model had salient factor loadings; however, the pattern was not stable - with three items loading on the *knowledge* factor and seven items on the *skills* factor.

The scale as currently used in the student survey is based primarily on a pragmatic approach where the NQF was used to guide the development of items based on the practicality and feasibility of their administration. As suggested in the literature, formulation of a LOs scale should be founded on a strong theory (Pike, 2011). Nusche's (2008) framework presents a summary of the key dimensions that need to be considered when assessing learning outcomes in higher education. As indicated by Nusche, LOs can be categorised into cognitive and non-cognitive. Based on this

categorization, LOs can be represented as unidimensional or multidimensional depending on the choice of indicators and the purpose of assessment. If multidimensionality of the scale is desired, it is reasonable that NOKUT revises the concept and generate more items to capture all dimensions in a more representative way and reconsider the scoring of the scale (McDonald, 1999). However, if the scale scores are to be used and interpreted as currently presented (sum scores compared across study programs), then each dimension applicable in the Norwegian setting should be developed and assessed separately.

Furthermore, literature recommends the use of generic transferable LOs when the assessment intends to compare LOs across disciplines (Caspersen, Smeby, et al., 2017; Gonyea, 2005; Nusche,2008). The modified scale capturing seven generic transferable skills (Nusche, 2008) fit the data well. This indicates that if the use of sum scores to compare study programs is desired, NOKUT may consider revising the scale to focus on items that capture the cognitive transferable skills' aspect of LO.

NOKUT's SSRLO measure compares students' satisfaction with their LOs across study programs. Measurement invariance test is a pre-requisite for justified study group comparison. The second research question thus entailed determining configural, metric and scalar invariance of the identified factor structure across four selected study programs. The results obtained showed that configural invariance was supported i.e., the single factor and factor loadings patterns for all four study groups are equivalent (Meredith, 1993). This indicates that the generic items in the SSRLO measure show the same structure in the four study groups. It can therefore be concluded that respondents from different study groups perceive the structure of the construct in the same way (Vandenberg & Lance 2000). The metric invariance test revealed equal factor loadings between the study programs. This shows that the measure has the same meaning and structure across the study groups – thus the relationship to the latent (LO) can be compared across the study groups (Steinmetz et al., 2009; Vandenberg & Lance 2000).

The results of the scalar invariance test show that the SSRLO measure is sensitive to study programs at the strong level – indicating that satisfaction with LOs as currently measured by the SSRLO does not have the same meaning across engineering, business and administration, teacher education, and nursing study programs. From this it can be concluded that the SSRLO scale does not have the same origin across the study programs. Accordingly, comparison and generalisation of claims based on comparisons of mean scores of these groups on the SSRLO cannot be made. However, the literature recommends that a bias study can be done to determine the specific items that function differently across the groups (Putnick & Bornstein, 2016; Vandenberg & Lance 2000). The support for partial scalar invariance reveals that for the same level of satisfaction with LOs, students across the four study groups will have different responses for item4 – '*Discipline and profession-specific skills'*, item 6 – '*Oral communication skills', and item 7 – 'cooperative skills'.* Observable and latent SSRLO scores are therefore only partially comparable across the four study programs. The non-invariance of these items can be interpreted based – to some extent – on the social organisation of knowledge networks within the disciplines and the areas of emphasis therein (Caspersen, Frølich, et al., 2017; Muller, 2009). Students self-select themselves to study programs that allow them to engage in and utilize their values and attitudes accordingly. Differences may therefore be evident in the answering of these SRLO items depending on the aspects of LO that the respective disciplines emphasize, individual differences, and how these environments socialise the students (Caspersen et al., 2014; Holland, 1997; Tourangeau et al., 2000). Future editing of the scale can be done considering these items if a valid comparison across the four study programs is desired.

This study makes contributions to existing literature. Methodologically, the study provides an example of a validation process with evidence based on the current use and interpretation of a SSRLO scale (Kane,2013) in the Norwegian context, in addition to empirical evidence on MI testing on self-reported scales (Vandenberg & Lance, 2000). Practically, the study contributes to the validation of an important national survey tool. The implication is, for valid use and interpretation of the scores from this scale, the indicators chosen should better reflect the desired factor structure,

and measurement equivalence across the study programs is vital. The study may to some extent justify the revision of the scale to better inform the interpretations and use of the scores by the users of the information. Future studies can develop a revised scale and validate it.

This study has some limitations. First, the learning outcomes scale in the student survey is a self-reported measure. Self-reported measures are commonly used in LO assessment and research in higher education. Yet compared to objective measures such as standardized test, student's subjective views of their learning may not always align with their actual mastery knowledge and skills due to e.g., the inability of some students to accurately report their LO and possible response bias. Nevertheless, SRLOs measures present advantages (e.g., their ease of administration, and the fact that assessment procedures needed are less demanding compared to standardized test) that make them popular in HE researches (Gonyea, 2005). Despite this, the limitations associated with self-reported data may affect the responses to the study items.

Second, the NOKUT categorization of study programs was used and only 4 study program types were selected. We recognise that there could be some differences even in these categories, for instance between engineering students. We aimed to test whether there were differences among the major academic groups as categorized by NOKUT, therefore for future studies, finer-grained analysis can be done.

In conclusion, NOKUT's SSRLO measure as currently presented does not reflect the intended interpretations and use of the scores. The implied unidimensionality is not supported and even with the readjustment of the factor model, some items were still found to be non-invariant across the studied groups, limiting group comparisons with these items in the scale. Nevertheless, these findings should be seen as one of the steps toward establishing the validity of the SSRLO.

References

Adam, S. (2006). *An introduction to learning outcomes*. Citeseer.

Adam, S. (2008). Learning outcomes based higher education: The Scottish experiences. *Bologna Seminar. Edinburgh*.

AERA, APA, & NCME. (2014). Standards for educational and psychological testing. *American Educational Research Association*.

Allan, J. (1996). Learning outcomes in higher education. *Studies in Higher Education*, *21*(1), 93–108.

Allen, J. P., & Van Der Velden, R. (2005). *The role of self-assessment in measuring skills*. ROA.

Bowman, N. A. (2011). Validity of college self-reported gains at diverse institutions. *Educational Researcher*, *40*(1), 22–24.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research*. Guilford publications.

Burrus, J., Jackson, T., Xi, N., & Steinberg, J. (2013). Identifying the most important 21st century workforce competencies: An analysis of the occupational information network(O*NET). *ETS Research Report Series*, *2013*(2), i–55. https://doi.org/10.1002/j.2333-8504.2013.tb02328.x

Caspersen, J., de Lange, T., Prøitz, T. S., Solbrekke, T. D., & Stensaker, B. (2011). LEARNING ABOUT QUALITY–. *Perspectives on Learning Outcomes and Their Operationalisations and Measurement*.

Caspersen, J., Frølich, N., Karlsen, H., & Aamodt, P. O. (2014). Learning outcomes across disciplines and professions: Measurement and interpretation. *Quality in Higher Education*, *20*(2), 195–215.

Caspersen, J., Frølich, N., & Muller, J. (2017). Higher education learning outcomes–Ambiguity and change in higher education. *European Journal of Education*, *52*(1), 8–19.

Caspersen, J., Smeby, J.-C., & Olaf Aamodt, P. (2017). Measuring learning outcomes. *European Journal of Education*, *52*(1), 20–30.

Chen, F. F. (2007). Sensitivity of Goodness of Fit Indexes to Lack of Measurement Invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(3), 464–504. https://doi.org/10.1080/10705510701301834

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating Goodness-of-Fit Indexes for Testing Measurement Invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(2), 233–255. https://doi.org/10.1207/S15328007SEM0902_5

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

Dias, D., & Soares, D. (2017). Learning outcomes in Higher Education: Designing a conceptual map for portuguese academia. *INTED2017 Proceedings*, 9188–9194.

Douglass, J. A., Thomson, G., & Zhao, C.-M. (2012). The learning outcomes race: The value of self-reported gains in large research universities. *Higher Education*, *64*(3), 317–335.

ECTS Users' Guide. (2005). *Brussels: Directorate-General for Education and Culture.* http://ec.europa.eu/education/programmes/socrates/ects/doc/guide_en.pdf

Gonyea, R. M. (2005). Self-reported data in institutional research: Review and recommendations. *New Directions for Institutional Research*, *2005*(127), 73–89.

Hansen, J. B., Gallavara, G., Nordblad, M., Persson, V., Salado-Rasmussen, J., & Weigelt, K. (2013). *Learning outcomes in external quality assurance approaches: Investigating and discussing Nordic practices and developments: Nordic quality assurance network in higher education*. Nordic Quality Assurance Network in Higher Education.

Holland, J. L. (1997). *Making vocational choices: A theory of vocational personalities and work environments*. Psychological Assessment Resources.

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., Rosseel, Y., Miller, P., Quick, C., Garnier-Villarreal, M., Selig, J., Boulton, A., & Preacher, K. (2020). semTools: Useful tools for structural equation modeling (0.5-3). *Computer Software]. Https://CRAN. R-Project. Org/Package= SemTools*.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73.

Kline, P. (2014). *An easy guide to factor analysis*. Routledge.

Kline, R. B. (2015). *Principles and practice of structural equation modeling*. Guilford publications.

Kraiger, K., Ford, J. K., & Salas, E. (1993). Application of cognitive, skill-based, and affective theories of learning outcomes to new methods of training evaluation. *Journal of Applied Psychology*, *78*(2), 311.

Kuh, G. D. (2005). *Putting student engagement results to use: Lessons from the field*.

Markle, R., Brenneman, M., Jackson, T., Burrus, J., & Robbins, S. (2013). Synthesizing frameworks of higher education student learning outcomes. *ETS Research Report Series*, *2013*(2), i–37.

Marsh, H. W. (1994). Confirmatory factor analysis models of factorial invariance: A multifaceted approach. *Structural Equation Modeling: A Multidisciplinary Journal*, *1*(1), 5–34.

McDonald, R. P. (1999). Test homogeneity, reliability, and generalizability. *Test Theory: A Unified Treatment*, 76–120.

McDonald, R. P. (2013). *Test theory: A unified treatment*. psychology press.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525–543. https://doi.org/10.1007/BF02294825

Muller, J. (2009). Forms of knowledge and curriculum coherence. *Journal of Education and Work*, *22*(3), 205–226. https://doi.org/10.1080/13639080902957905

NOKUT. (2019). *Studiebarometeret. Http://studiebarometeret.no/no/artikkel/2*.

Nusche, D. (2008). *Assessment of Learning Outcomes in Higher Education: A comparative review of selected practices* (OECD Education Working Papers No. 15; OECD Education Working Papers, Vol. 15). https://doi.org/10.1787/244257272573

Oswald, F. L., Schmitt, N., Kim, B. H., Ramsay, L. J., & Gillespie, M. A. (2004). Developing a biodata measure and situational judgment inventory as predictors of college student performance. *Journal of Applied Psychology*, *89*(2), 187.

Ouimet, J. A., Bunnage, J. C., Carini, R. M., Kuh, G. D., & Kennedy, J. (2004). Using focus groups, expert advice, and cognitive interviews to establish the validity of a college student survey. *Research in Higher Education*, *45*(3), 233–250.

Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research*, *74*(4), 525–556.

Pike, G. R. (1992). Using mixed-effect structural equation models to study student academic development. *The Review of Higher Education*, *15*(2), 151–177.

Pike, G. R. (1993). The relationship between perceived learning and satisfaction with college: An alternative view. *Research in Higher Education*, *34*(1), 23–40.

Pike, G. R. (1996). Limitations of using students' self-reports of academic development as proxies for traditional achievement measures. *Research in Higher Education*, *37*(1), 89–114.

Pike, G. R. (2011). Using college students' self-reported learning outcomes in scholarly research. *New Directions for Institutional Research*, *2011*(150), 41–58.

Pike, G. R., Smart, J. C., & Ethington, C. A. (2012). The mediating effects of student engagement on the relationships between academic disciplines and learning outcomes: An extension of Holland's theory. *Research in Higher Education*, *53*, 550–575.

Porter, S. R. (2013). Self-reported learning gains: A theory and test of college student survey response. *Research in Higher Education*, *54*(2), 201–226.

Prøitz, T. S. (2010). Learning outcomes: What are they? Who defines them? When and where are they defined? *Educational Assessment, Evaluation and Accountability*, *22*(2), 119–137.

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review*, *41*, 71–90. https://doi.org/10.1016/j.dr.2016.06.004

R. Core Team. (2020). R: A language and environment for statistical computing. (4.0. 3). R Foundation for Statistical Computing. *URL Http://Www*.

Revelle, W. R. (2017). *psych: Procedures for personality and psychological research*.

Revelle, W., & Zinbarg, R. E. (2009). Coefficients Alpha, Beta, Omega, and the glb: Comments on Sijtsma. *Psychometrika*, *74*(1), 145–154. https://doi.org/10.1007/s11336-008-9102-z

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, *17*(3), 354–373. https://doi.org/10.1037/a0029315

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*, 1–36.

Rutkowski, L., & Svetina, D. (2014). Assessing the Hypothesis of Measurement Invariance in the Context of Large-Scale International Surveys. *Educational and Psychological Measurement*, *74*(1), 31–57. https://doi.org/10.1177/0013164413498257

Steinmetz, H., Schmidt, P., Tina-Booh, A., Wieczorek, S., & Schwartz, S. H. (2009). Testing measurement invariance using multigroup CFA: Differences between educational groups in human values measurement. *Quality & Quantity*, *43*(4), 599–616.

Stensaker, B., & Sweetman, R. (2014). Impact of assessment initiatives on quality assurance. *Higher Education Learning Outcomes Assessment*, 237–259.

Tabachnick, B. G., Fidell, L. S., & Ullman, J. B. (2007). *Using multivariate statistics* (Vol. 5). pearson Boston, MA.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). The psychology of survey response. *Cambridge and New York: Cambridge University Press*.

Tremblay, K., Lalancette, D., & Roseveare, D. (2012). Assessment of Higher Education Learning Outcomes: Feasibility Study Report, Volume 1–Design and Implementation. *Paris, France: Organisation for Economic Co-Operation and Development*, *1*.

Van de Vijver, F., & Leung, K. (1997). Methods and data analysis of comparative research. *Handbook of Cross-Cultural Psychology*, *1*, 257–300.

Van de Vijver, F., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology*, *54*(2), 119–135. https://doi.org/10.1016/j.erap.2003.12.004

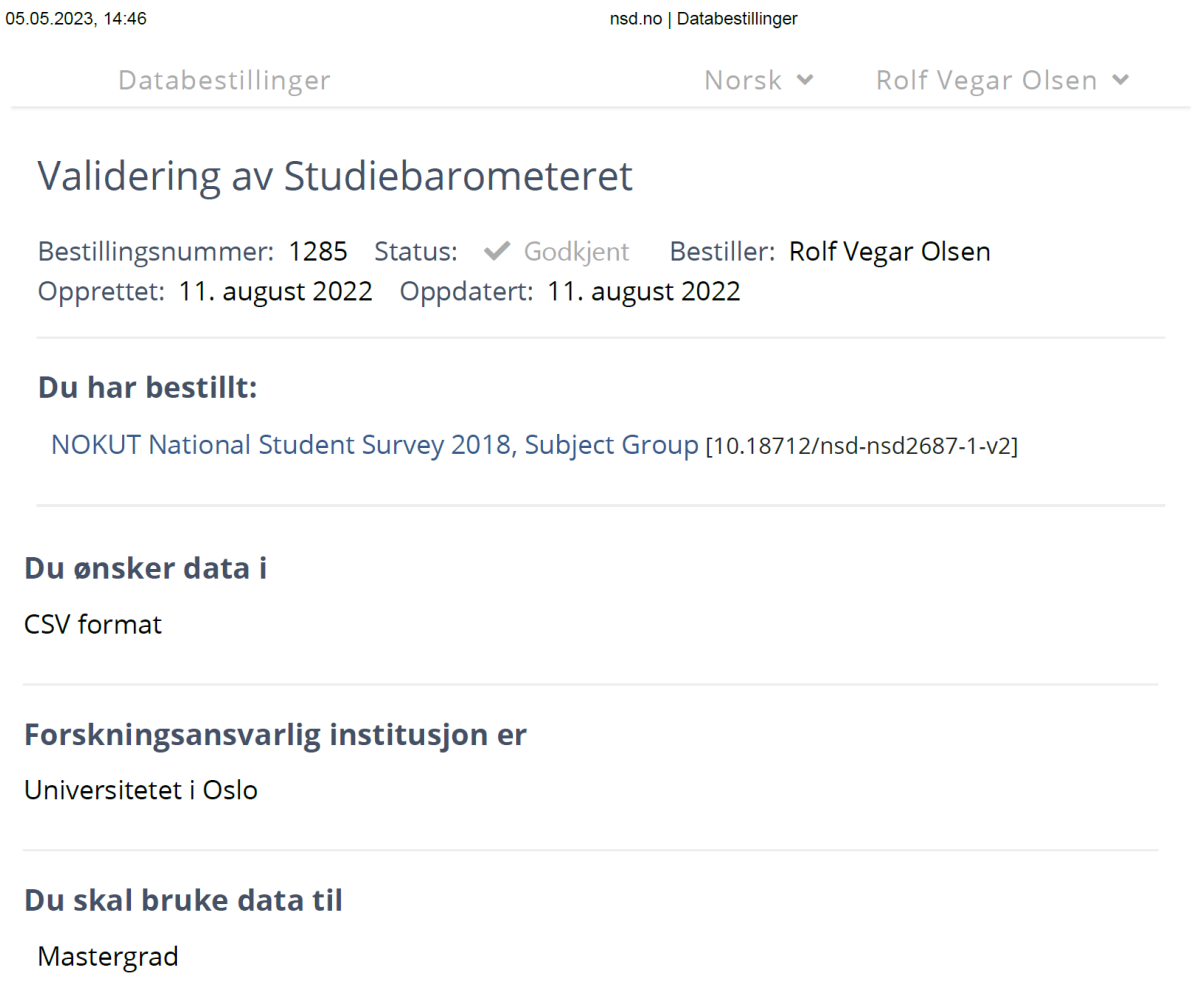
Vandenberg, R. J., & Lance, C. E. (2000). A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research. *Organizational Research Methods*, *3*(1), 4–70. https://doi.org/10.1177/109442810031002

Appendix I: GDPR documents & Ethical approval

Ethical approval

We applied and obtained approval from the Norwegian Agency for Shared Services in Education and Research (SIKT) – formerly Norwegian centre for research data (NSD). Our study did not involve processing of personal data therefore, after signing the agreement, data was made available to us through the SIKT web page. No further ethical approval was needed because an agreement was signed by the respondents before taking part in the survey– approving that anonymous data can be used for research purposes. Copies of SIKT application and approval, and signed agreement are presented below.

NSD Application



05.05.2023, 14:46 nsd.no | Databestillinger

Databestillinger Norsk Rolf Vegar Olsen

Validering av Studiebarometeret

Bestillingsnummer: 1285 Status: ✔ Godkjent Bestiller: Rolf Vegar Olsen
Opprettet: 11. august 2022 Oppdatert: 11. august 2022

Du har bestillt:

NOKUT National Student Survey 2018, Subject Group [10.18712/nsd-nsd2687-1-v2]

Du ønsker data i

CSV format

Forskningsansvarlig institusjon er

Universitetet i Oslo

Du skal bruke data til

Mastergrad

Veileder på studentoppgave

rolfvo@cemo.uio.no Fjern

Hvem skal ha tilgang til data?

rolfvo@uio.no Fjern

a.r.lorenz@imv.uio.no Fjern

lucywg@student.uv.uio.no Fjern

Mottakers epost ved institusjonen

Hendelseslogg

Opprettet 11. august 2022 - 09:41:06

Databestillinger Norsk Rolf Vegar Olsen

Innlevert 11. august 2022 - 09:42:16

R

Rolf Vegar Olsen 11. august 2022 - 09:43:13

Jeg antar at data som dere gjør tilgjengelig på denne måten ikke trenger ytterligere meldinger til NSD og vårt eget personvernombud?

Cecilie Hopland Jentoft 11. august 2022 - 11:53:36

Hei Vi har fire datasett fra studiebarometeret 2018 for bestilling. Alla datasettene med unntak av Studiebarometeret 2018, grunnfil er uten personidentifiserende

opplysninger og gjøres tilgjengelig, uten at det er behov for melding til Sikt (tidligere NSD) og/eller eget personvernombud. Du har bestilt datasettet Studiebarometeret 2018, faggruppe og for at det skal gjøres tilgjengelig må du legge til studentenes e-postadresse, så vil vi sende ut avtal om bruk av data til dem og en avtale til de som veilede. Nå avtalen er signert, vil dataene bli tilgjengelig for nedlasting. Med vennlig hilsen Cecilie Hopland Jentoft

R **Rolf Vegar Olsen** 11. august 2022 - 11:58:07

Takk. Da skal alt være i god orden her

Godkjent 11. august 2022 - 12:24:38

Databestillinger Norsk Rolf Vegar Olsen

Send melding

Signed Agreement for Data Access

Dokumentet er signert digitalt av følgende undertegnere:

Det signerte dokumentet innholder
• En forside med informasjon om signaturene
• Alle originaldokumenter med signaturer på hver side
• Digitale signaturer

Dokumentet er forseglet av Posten Norge
Signeringen er gjort med digital signering levert av Posten Norge AS. Posten garanterer for autentisiteten og forseglingen av dette dokumentet.

Slik ser du at signaturene er gyldig
Hvis du åpner dette dokumentet i Adobe Reader, skal det stå øverst at dokumentet er sertifisert av Posten Norge AS. Dette garanterer at innholdet i dokumentet er ikke endret etter signering.

- LUCY WAIRIMU GITIRIA, signert 15.08.2022 med ID-Porten: BankID Mobil

Sikt
Kunnskapssektorens tjenesteleverandør

Information in English

User Agreement for Data Access

You are hereby granted access to use the mentioned dataset in the project "*Validering av Studiebarometeret",* as described in order number 1285. Access to the data will be granted once the enclosed agreement has been signed.

As a user, you commit yourself to:

1. Only use the data for the project described in the application.
   If you want to use the data for another purpose, you must send a new application.
2. Not give others access to the dataset.

If others assist you in the use of the data, they too must sign an agreement.

3. Delete the data file(s), or to apply for an extended deadline for data deletion after the projects ends, or at the *latest by August 11, 2024*
4. Not attempt to identify any individuals in the dataset.

   If you were to identify an individual, you must notify Sikt.
5. Cite the producer and distributor of the data.

I am aware that researcher's duty of secrecy is regulated by the Public Administration Act § 13e. I am also aware that an intentional or negligent breach of the duty of secrecy, or complecity in this, can be punished by fines or imprisonment.

This user agreement concerns the following data

- **NOKUT National Student Survey 2018, Subject Group [10.18712/nsd-nsd2687-1-v2]**

Sikt — Kunnskapssektorens tjenesteleverandør | Norwegian Agency for Shared Services in Education and Research
Tel: (+47) 73 98 40 40 | postmottak@sikt.no | Org.nr: 919 477 822

Besøksadresser: Trondheim: Abels gate 5 — Teknobyen • Oslo: Fridtjof Nansens vei 19 • Bergen: Harald Hårfagres gate 29 www.sikt.no

Appendix II: Data Management and Analysis Code

```
## PACKAGES NEEDED
#############################################################################
library(readxl)
library(lavaan)
library(psych)
library(tidyverse)
library(GPArotation)
library(corrplot)
library(semTools) # for additional functions in SEM
library(semPlot) # for path diagram
library(effsize) # for calculating effect size
library(rcompanion)

#**DATA*
#############################################################################
#Load data
getwd()
setwd("D:/MA ED Measurement Evaluation and Assessment")
NSD2687_1_no_1 <- read_excel("THESIS 2022/R scripts/NSD2687-1-no 1.xlsx")
#View(NSD2687_1_no_1)

#DATA PREPARATION
Data<-NSD2687_1_no_1
preli.df<-Data[c(6,7,13,14,93:102)]
apply(preli.df[1:14],2,table,exclude=NULL)#A glance at the data
preli.df[1:14][preli.df[1:14]==9999]<-NA # change 9999 to NAs
preli.df[1:14][preli.df[1:14]==999]<-NA # change 999 to NAs

#Missing Data
missingdata<-preli.df
business<-missingdata[missingdata$Utd_type == 'Ã˜KADM',]
enginers<-missingdata[missingdata$Utd_type == 'INGENIÃ˜R',]  #Engineering
MAnurses<-missingdata[missingdata$Utd_type == 'SYKEPLEIE-MA',]#Nursing
Masters
BAnurses<-missingdata[missingdata$Utd_type == 'SYKEPLEIE',] # Nursing-
bachelors
Teachers<-missingdata[missingdata$Utd_type == 'GRUNNSKOLE',]

#Merge MA and BA nurses
listnurses<- list(MAnurses,BAnurses)
allnurses<-Reduce(function(x, y) merge(x, y, all=TRUE), listnurses)

#change rows names
business$Utd_type[business$Utd_type == 'Ã˜KADM'] <- 'AKADM'
enginers$Utd_type[enginers$Utd_type == 'INGENIÃ˜R'] <- 'ENGINER'
allnurses$Utd_type[allnurses$Utd_type == 'SYKEPLEIE-MA'] <- 'SYKEPLEIE'

#full set
listfullset<- list(business,enginers,allnurses,Teachers)
dfmissingness<-Reduce(function(x, y) merge(x, y, all=TRUE), listfullset)

#**Check percentage of missingness*
#Check columns with > 5%- use a function and apply
percent.miss<-function(x){sum(is.na(x))/length(x)*100} #finds the nas,sums
them
#divided by total and multiplied by 100 to give %
names(dfmissingness)

#Check column in each group
```

```
apply(business[,-c(1,2,3,4)],2,percent.miss)#without the categorical
variables
apply(enginers[,-c(1,2,3,4)],2,percent.miss)#without the categorical
variables
apply(allnurses[,-c(1,2,3,4)],2,percent.miss)#without the categorical
variables
apply(Teachers[,-c(1,2,3,4)],2,percent.miss)#without the categorical
variables
#Percentage of those missing some data points per group
(sum(is.na(business[c(5:14)]))/prod(dim(business[c(5:14)])))*100
(sum(is.na(enginers[c(5:14)]))/prod(dim(enginers[c(5:14)])))*100
(sum(is.na(allnurses[c(5:14)]))/prod(dim(allnurses[c(5:14)])))*100
(sum(is.na(Teachers[c(5:14)]))/prod(dim(Teachers[c(5:14)])))*100

#use apply to check for columns whole dataset
apply(dfmissingness[,-c(1,2,3,4)],2,percent.miss)#without the categorical
variables
#**All above 5% cannot impute. Use listwise deletion to have uniform subset
across all analysis (Peugh&Ender,2004)*

#check rows
missing.in.rows<-apply(dfmissingness[,-c(1,2,3,4)],1,percent.miss)
table(missing.in.rows)#

#Percentage of those missing some data points- whole dataset
(sum(is.na(dfmissingness[c(5:14)]))/prod(dim(dfmissingness[c(5:14)])))*100

#Remove NAs-removes all rows with atleast 1 value missing (remains with
only complete cases)
preli.df1<-preli.df[complete.cases(preli.df[5:14]),]
apply(preli.df1[1:14],2,table,exclude=NULL)#A glance at the subset data key
variables=7488

#  SUBSET 4 STUDY PROGRAMS
dfAKADM<-preli.df1[preli.df1$Utd_type == 'Ã˜KADM',]           # business
and administration-Enterprising /Ã˜KADM/
dfENGINEERING<-preli.df1[preli.df1$Utd_type == 'INGENIÃ˜R',]  #Engineering
dfSYKEP_MA<-preli.df1[preli.df1$Utd_type == 'SYKEPLEIE-MA',]  #Nursing
Masters
dfSYK.BA<-preli.df1[preli.df1$Utd_type == 'SYKEPLEIE',]       # Nursing-
bachelors
dfGRUNNSKOLE<-preli.df1[preli.df1$Utd_type == 'GRUNNSKOLE',]  # Primary
teacher education

#merge Nursing Masters and Nursing bachelors
listSYKEP<- list(dfSYKEP_MA,dfSYK.BA)
dfSYKEPLEIE<-Reduce(function(x, y) merge(x, y, all=TRUE), listSYKEP)

#A glance at the subset data
apply(dfGRUNNSKOLE[1:14],2,table,exclude=NULL)
apply(dfSYKEPLEIE[1:14],2,table,exclude=NULL)
apply(dfAKADM[1:14],2,table,exclude=NULL)
apply(dfENGINEERING[1:14],2,table,exclude=NULL)

#Change Row names for 2 programs
dfAKADM$Utd_type[dfAKADM$Utd_type == 'Ã˜KADM'] <- 'AKADM'
dfENGINEERING$Utd_type[dfENGINEERING$Utd_type == 'INGENIÃ˜R'] <- 'ENGINER'
dfSYKEPLEIE$Utd_type[dfSYKEPLEIE$Utd_type == 'SYKEPLEIE-MA'] <- 'SYKEPLEIE'

#DATA FOR ANALYSIS
#Make list of the four groups data-sets
```

```
list0<- list(dfGRUNNSKOLE,dfSYKEPLEIE,dfAKADM,dfENGINEERING)
#Merge all data frames in list
df1<-Reduce(function(x, y) merge(x, y, all=TRUE), list0)

#Change column names
df2<-df1
colnames(df2)
colnames(df2) <- c("Study_Program","Study_Group", "Gender", "YearofStudy",
                   "Item1","Item2","Item3","Item4","Item5","Item6","Item7",
                    "Item8","Item9","Item10")

df3<-df2[c(1,5:14)] # MGCFA- Study program Analysis data.

apply(df3[1:11],2,table,exclude=NULL)#A glance at the subset data

##############################################################################
#**DATA SUITABILITY CHECK**
##############################################################################
#**Univariate descriptive statistics**
describe(df3[, 2:11])

#**% of response to options per item*,
response.frequencies(df3[,2:11]) #All response options are used, and there
are no missing values.

#**Linearity check*
pairwiseCount(df3[,2:11])

#**Bivariate characteristics of the data* #psych package (Revelle, 2019)
pairs.panels(df3[,2:11], stars = TRUE)
#correlations moderate and pretty close-except item 2 and 3, not normal,
linear relation

#**Multivariate descriptive statistics**
mardia(df3[, 2:11]) #multivariate skweness and kurtosis

#Plot multivariate normality histogram
#library(rcompanion)
plotNormalHistogram(df3[, 2:11],prob = FALSE)

#**Covariance and correlation matrices*
#Covariance
cov.mx <- round(cov(df3[, 2:11], method = "pearson"), digits = 2)
cov.mx[upper.tri(cov.mx)] <- NA # redundant information set to NA, we will
only get the lower triangle of the matrix
print(cov.mx, na.print = "")

#**multivariate outliers*
#identify the outlier values
mahl<-mahalanobis(df3[, -c(1)],colMeans(df3[,-c(1)]),cov(df3[, -c(1)],use =
"pairwise.complete"))

#summary of the outliers values
mahl
summary(mahl) #min outlier value is 1.031 and max is 74.85

#calculate cutoff score for p<0.001
cutoff<-qchisq(1-.001,ncol(df3[,-c(1)]))
cutoff        #any values above the cuttoff value of 18.307- are considered
multivariate outliers
```

```
#Identify number of participants who exceed the cutoff score
summary(mahl<cutoff)
#205 participants are multivariate outliers

#Decide whether to keep or remove value
#if remove
df.without_outliers=df3[mahl<cutoff,]

#Keep and do sensitivity analysis*

#**Use cov2cor() to convert the covariance matrix into a correlation
matrix*
corr.mx <- round(cov2cor(cov.mx), digits = 2)
corr.mx[upper.tri(corr.mx)] <- NA
print(corr.mx, na.print = "")

#**visual on the magnitudes of correlations**
#Heatmap
cor_tab <- cor(df3[, 2:11], use = "pairwise.complete.obs")
cor_tab
corPlot(cor_tab, numbers = TRUE)

#Reliability
psych::alpha(df3[, 2:11]) #.88

#############################################################################
#**CONFIRMATORY FACTOR ANALYSIS**
#############################################################################
#**Conceptual model**

cfa.ModelA<- " Skills Achievement =~ Item4+Item5+Item6+Item7+
                                     Item8+Item9+Item10
                       Knowledge Achievement =~ Item1+Item2+Item3"


cfa.ModelA.fit = lavaan::cfa(cfa.ModelA, data = df3[(2:11)], estimator =
"MLR")

summary(cfa.ModelA.fit, fit.measures = T, standardized = T,rsquare=T)
fitMeasures(cfa.ModelA.fit,
c("cfi.scaled","tli.scaled","rmsea.scaled","srmr"))

#**Locale fit**
lavResiduals(cfa.ModelA.fit)# "cor.bentler" table-cases of above 0.1

modindices(cfa.ModelA.fit, sort = TRUE, maximum.number = 5)# since we are
using MLR, look at 'mi'
#suggested model improvements Item2~~Item3

cfa.ModelAi<- " Skills Achievement =~ Item4+Item5+Item6+Item7+
                                     Item8+Item9+Item10
                       Knowledge Achievement =~ Item1+Item2+Item3
                       #Covariance
                       Item2~~Item3"
cfa.ModelAi.fit = lavaan::cfa(cfa.ModelAi, data = df3[(2:11)], estimator =
"MLR")

summary(cfa.ModelAi.fit, fit.measures = T, standardized = T,rsquare=T)

#**Plot the path Diagram**
semPaths(cfa.ModelAi.fit, 'path', 'std', style = 'lisrel',
```

```
         edge.color = 'black', intercepts = F)

##############################################################################
#**Unidimensional model with all items**

cfa.ModelB <- "Learning_Outcomes
=~Item1+Item2+Item3+Item4+Item5+Item6+Item7+
                                    Item8+Item9+Item10"

cfa.ModelB.fit = lavaan::cfa(cfa.ModelB, data = df3[(2:11)], estimator =
"MLR")

summary(cfa.ModelB.fit, fit.measures = T, standardized = T,rsquare=T)
fitMeasures(cfa.ModelB.fit,c("cfi.scaled","tli.scaled","rmsea.scaled","srmr
"))
#**Poor fit**
##############################################################################
#**Reduced scale**- focus on NOKUT's intention of measuring generic skills
# Revised Measurement Model- with focus on the generic skills

cfa.ModelC <- "Learning_Outcomes =~Item4+Item5+Item6+Item7+
                                    Item8+Item9+Item10"

cfa.ModelC.fit <- cfa(cfa.ModelC, data = df3[c(5:11)], estimator= "MLR")

fitMeasures(cfa.ModelC.fit,
c("cfi.scaled","tli.scaled","rmsea.scaled","srmr"))

summary(cfa.ModelC.fit, fit.measures = T, standardized = T,rsquare=T)

#**Plot the path Diagram**
semPaths(cfa.ModelC.fit, 'path', 'std', style = 'lisrel',
         edge.color = 'black', intercepts = F)

#Reliability of the reduced scale
psych::alpha(df3[, 5:11]) #.0.86
omegaSem(df3[c(5:11)],1) #the confirmatory solution #.84

##############################################################################
#**MEASUREMENT INVARIANCE**
##############################################################################
##**Run CFA separately in each group.**
# Reduced scale
df4<-df3[c(1,5:11)]
str(df4)
table(df4$Study_Program)# check out frequency of variable of interest
#Business and Admin
modelfit.B_A <- cfa(cfa.ModelC, data = df4[df4[, 1] == "AKADM", ],
estimator= "MLR")

summary(modelfit.B_A, fit.measures = T, standardized = T,rsquare=T)

fitMeasures(modelfit.B_A,
c("cfi.scaled","tli.scaled","rmsea.scaled","srmr"))

#**Local misfit*
lavResiduals(modelfit.B_A)
modindices(modelfit.B_A, sort = TRUE, maximum.number = 5)

#**Plot the path Diagram**
semPaths(modelfit.B_A, 'path', 'std', style = 'lisrel',
```

```
         edge.color = 'black', intercepts = F)
#########################################################################
#Engineering
modelfit.E <- cfa(cfa.ModelC, data = df4[df4[, 1] == "ENGINER", ],
estimator= "MLR")
summary(modelfit.E, fit.measures = T, standardized = T,rsquare=T)
fitMeasures(modelfit.E, c("cfi.scaled","tli.scaled","rmsea.scaled","srmr"))

#**Local misfit*
lavResiduals(modelfit.E)
modindices(modelfit.E, sort = TRUE, maximum.number = 5)

#**Plot the path Diagram**
semPaths(modelfit.E, 'path', 'std', style = 'lisrel',
         edge.color = 'black', intercepts = F)
#########################################################################
#Teacher education
modelfit.G <- cfa(cfa.ModelC, data = df4[df4[, 1] == "GRUNNSKOLE", ],
estimator= "MLR")
summary(modelfit.G, fit.measures = T, standardized = T,rsquare=T)
fitMeasures(modelfit.G, c("cfi.scaled","tli.scaled","rmsea.scaled","srmr"))

#**Local misfit*
lavResiduals(modelfit.G)
modindices(modelfit.G, sort = TRUE, maximum.number = 5)

#**Plot the path Diagram**
semPaths(modelfit.G, 'path', 'std', style = 'lisrel',
         edge.color = 'black', intercepts = F)
#########################################################################
#Nursing
modelfit.N <- cfa(cfa.ModelC, data = df4[df4[, 1] == "SYKEPLEIE", ],
estimator= "MLR")
summary(modelfit.N, fit.measures = T, standardized = T,rsquare=T)
fitMeasures(modelfit.N, c("cfi.scaled","tli.scaled","rmsea.scaled","srmr"))

#**Local misfit*
lavResiduals(modelfit.N)
modindices(modelfit.N, sort = TRUE, maximum.number = 5)

#**Plot the path Diagram**
semPaths(modelfit.N, 'path', 'std', style = 'lisrel',
         edge.color = 'black', intercepts = F)
#########################################################################
#**STEP 1. Configural invariance: Equal form**
#########################################################################
#**Fit the baseline model for all groups simultaneously**
fit.MI1 <- cfa(cfa.ModelC, data = df4, estimator = "MLR", group =
"Study_Program",std.lv=TRUE, meanstructure = TRUE)
summary(fit.MI1, standardized = TRUE, fit.measures = TRUE)
fitMeasures(fit.MI1, c("cfi.scaled","rmsea.scaled","srmr"))
#########################################################################
#**STEP 2. Metric invariance: Equal factor loadings**
#########################################################################
fit.MI2 <- cfa(cfa.ModelC, data = df4, estimator = "MLR", group =
"Study_Program",
               std.lv=TRUE, meanstructure = TRUE, group.equal =
c("loadings"))
summary(fit.MI2, standardized = TRUE, fit.measures = TRUE)
fitMeasures(fit.MI2, c("cfi.scaled","rmsea.scaled","srmr"))
```

```
#**Compare the model fit indices**
overallfit0<-compareFit(fit.MI2, fit.MI1)
summary(overallfit0,fit.measures = c("cfi.scaled","rmsea.scaled","srmr"))
##############################################################################
#**STEP 3. Scalar invariance: Equal factor loadings and intercepts**
##############################################################################
fit.MI3 <- cfa(cfa.ModelC, data = df4, estimator = "MLR", group =
"Study_Program",
             std.lv=TRUE,  meanstructure = TRUE, group.equal =
c("loadings", "intercepts"))
summary(fit.MI3, standardized = TRUE, fit.measures = TRUE)
fitMeasures(fit.MI3, c("cfi.scaled","rmsea.scaled","srmr"))

#**Compare models**
overallfit1<-compareFit(fit.MI3, fit.MI2) #from SemTools
summary(overallfit1,fit.measures = c("cfi.scaled","rmsea.scaled","srmr"))
##############################################################################
#**Partial invariance**
##############################################################################
#list of models estimated
models<-
list(fit.configural=fit.MI1,fit.loadings=fit.MI2,fit.intercepts=fit.MI3,ref
group=1)
partialInvariance(models,type = "scalar")
#**Proposes the item to free i.e., item 4.

#**Adjust the model releasing those items one at a time and compare model
to metric model*
#*item 4 released
fit.MI3a <- cfa(model=cfa.ModelC,
                data = df4,
                estimator = "MLR",
                group = "Study_Program",
                meanstructure = TRUE,
                std.lv=TRUE,
                group.equal = c("loadings", "intercepts"),
                group.partial=c("Item4 ~1"))

summary(fit.MI3a, standardized = TRUE, fit.measures = TRUE)
fitMeasures(fit.MI3a, c("cfi.scaled","rmsea.scaled","srmr"))

#**Compare models**
overallfit2<-compareFit(fit.MI3a, fit.MI2) #from SemTools
summary(overallfit2,fit.measures = c("cfi.scaled","rmsea.scaled","srmr"))
#**Not yet*
##############################################################################
#list of models estimated
models1<-
list(fit.configural=fit.MI1,fit.loadings=fit.MI2,fit.intercepts=fit.MI3a,re
fgroup=1)
partialInvariance(models1,type = "scalar")
#Proposes item 7

#**Adjust the model again* item7 released in addition to item 4
fit.MI3b <- cfa(model=cfa.ModelC,
                data = df4,
                estimator = "MLR",
                group = "Study_Program",
                meanstructure = TRUE,
                std.lv=TRUE,
                group.equal = c("loadings", "intercepts"),
```

```
                group.partial=c("Item4 ~1","Item7 ~1"))

summary(fit.MI3b, standardized = TRUE, fit.measures = TRUE)
fitMeasures(fit.MI3b, c("cfi.scaled","rmsea.scaled","srmr"))

#**Compare models**
overallfit3<-compareFit(fit.MI3b, fit.MI2) #from SemTools
summary(overallfit3,fit.measures = c("cfi.scaled","rmsea.scaled","srmr"))
#**Not yet*
##############################################################################
#list of models estimated
models2<-
list(fit.configural=fit.MI1,fit.loadings=fit.MI2,fit.intercepts=fit.MI3b,re
fgroup=1)
partialInvariance(models2,type = "scalar")
#Proposes item 6

#**Adjust the model again* item6 released in addition to item 4 & 7
fit.MI3c <- cfa(model=cfa.ModelC,
                data = df4,
                estimator = "MLR",
                group = "Study_Program",
                meanstructure = TRUE,
                std.lv=TRUE,
                group.equal = c("loadings", "intercepts"),
                group.partial=c("Item4 ~1","Item7 ~1","Item6 ~1"))

summary(fit.MI3c, standardized = TRUE, fit.measures = TRUE)
fitMeasures(fit.MI3c, c("cfi.scaled","rmsea.scaled","srmr"))

#**Compare models**
overallfit4<-compareFit(fit.MI3c, fit.MI2) #from SemTools
summary(overallfit4,fit.measures = c("cfi.scaled","rmsea.scaled","srmr"))

#** partial scalar invariance achieved after freely estimating the
intercepts of 3 items*
#**(Item4, item7 and item6)*
##############################################################################
#**Compare latent means across groups putting into consideration the non-
invariant intercepts*
##############################################################################
#**Pull parameter estimates from the last model*
para.ests<-parameterestimates(fit.MI3c)
para.ests
###################################
#**For each group- Extract unstandardized factor loadings(=~)*'est' is the
column with factor loading
AKAM_load<-subset(para.ests, group=='1' & op %in% '=~',select="est")
AKAM_load
###################################
ENGI_load<-subset(para.ests, group=='2' & op %in% '=~',select="est")
ENGI_load
###################################
GRUN_load<-subset(para.ests, group=='3' & op %in% '=~',select="est")
GRUN_load
###################################
SYKP_load<-subset(para.ests, group=='4' & op %in% '=~',select="est")
SYKP_load
###################################
#**For each group- Extract intercept (~1)*'est' is the column with latent
means
```

```
###################################
#Extract intercepts for each group and remove the last row
AKADintercepts<-subset(para.ests, group=='1' & op %in% '~1',select="est")
AKADintercepts
AKAd.intercepts<-AKADintercepts[-c(8),]#Remove last row because its the
intercept for the factor
AKAd.intercepts
###################################
ENGI_intercepts<-subset(para.ests, group=='2' & op %in% '~1',select="est")
ENGI_intercepts
ENGI_intercepts<-ENGI_intercepts[-c(8),]#Remove last row because its the
intercept for the factor
ENGI_intercepts
###################################
GRUN_intercepts<-subset(para.ests, group=='3' & op %in% '~1',select="est")
GRUN_intercepts
GRUN_intercepts<-GRUN_intercepts[-c(8),]#Remove last row because its the
intercept for the factor
GRUN_intercepts
###################################
SYKP_intercepts<-subset(para.ests, group=='4' & op %in% '~1',select="est")
SYKP_intercepts
SYKP_intercepts<-SYKP_intercepts[-c(8),] #Remove last row because its the
intercept for the factor
SYKP_intercepts
###################################
#**Figure out the Means* #minus the intercepts from a persons observed
score and divide the result by unstandardized factor loadings
###################################
AKADoperation<-function(x){((x-AKAd.intercepts)/AKAM_load)} # X-person's
observed score
AKADfactorintercept<-apply(df4[,c(2:8)],1,AKADoperation)#apply the group
operation, on rows(1),on the LO items.
AKADfactorintercept
###################################
ENGIoperation<-function(x){((x-ENGI_intercepts)/ENGI_load)}
ENGIfactorintercept<-apply(df4[,c(2:8)],1,ENGIoperation)#apply the group
operation, on rows(1),on the LO items
ENGIfactorintercept
###################################
GRUNoperation<-function(x){((x-GRUN_intercepts)/GRUN_load)}
GRUNfactorintercept<-apply(df4[,c(2:8)],1,GRUNoperation)#apply the group
operation, on rows(1),on the LO items
GRUNfactorintercept
###################################
SYKPoperation<-function(x){((x-SYKP_intercepts)/SYKP_load)}
SYKPfactorintercept<-apply(df4[,c(2:8)],1,SYKPoperation)#apply the group
operation, on rows(1),on the LO items
SYKPfactorintercept
#######################################
#**Save the lists created as data-frame*
#######################################
dfAKADfactorintercept<-as.data.frame(AKADfactorintercept)
dfENGIfactorintercept<-as.data.frame(ENGIfactorintercept)
dfGRUNfactorintercept<-as.data.frame(GRUNfactorintercept)
dfSYKPfactorintercept<-as.data.frame(SYKPfactorintercept)
dfSYKPfactorintercept
#############################################################################
#**Flip the data-frames to have the respondents per row using the 't'
function*
#############################################################################
```

```
AKADfactorinterceptdf<-t(dfAKADfactorintercept)
ENGIfactorinterceptdf<-t(dfENGIfactorintercept)
GRUNfactorinterceptdf<-t(dfGRUNfactorintercept)
SYKPfactorinterceptdf<-t(dfSYKPfactorintercept)
########################################
#**Save the matrix created as data-frame*
########################################
Afact.inter<-as.data.frame(AKADfactorinterceptdf)
Efact.inter<-as.data.frame(ENGIfactorinterceptdf)
Gfact.inter<-as.data.frame(GRUNfactorinterceptdf)
Sfact.inter<-as.data.frame(SYKPfactorinterceptdf)
#########################################################################
#**Calculate the row mean for each respondent*
#########################################################################
AKADlatentmeans<-rowMeans(Afact.inter)
ENGIlatentmeans<-rowMeans(Efact.inter)
GRUNlatentmeans<-rowMeans(Gfact.inter)
SYKPlatentmeans<-rowMeans(Sfact.inter)
View(ENGIlatentmeans)

SDAKADM<-as.data.frame(AKADlatentmeans)
SDENGI<-as.data.frame(ENGIlatentmeans)
SDGRUN<-as.data.frame(GRUNlatentmeans)
SDSYKP<-as.data.frame(SYKPlatentmeans)
View(SDAKADM)
#########################################################################
#**Calculate t-test*
#########################################################################
#group 1 and 2
t.test(AKADlatentmeans,ENGIlatentmeans,
       alternative = "two.sided",
       paired = FALSE,
       var.equal = TRUE, #Welch t-test
       na.action=TRUE,
       p.adjust.method = "bonferroni")

#group 1 and 3
t.test(AKADlatentmeans,GRUNlatentmeans,
       alternative = "two.sided",
       paired = FALSE,
       var.equal = FALSE,
       na.action=TRUE,
       p.adjust.method = "bonferroni")

#group 1 and 4
t.test(AKADlatentmeans,SYKPlatentmeans,
       alternative = "two.sided",
       paired = FALSE,
       var.equal = FALSE,
       na.action=TRUE,
       p.adjust.method = "bonferroni")

#group 2 and 3
t.test(ENGIlatentmeans,GRUNlatentmeans,
       alternative = "two.sided",
       paired = FALSE,
       var.equal = FALSE,
       na.action=TRUE)

#group 2 and 4
t.test(ENGIlatentmeans,SYKPlatentmeans,
```

```
       alternative = "two.sided",
       paired = FALSE,
       var.equal = FALSE,
       na.action=TRUE)

#group 3 and 4
t.test(GRUNlatentmeans,SYKPlatentmeans,
       alternative = "two.sided",
       paired = FALSE,
       var.equal = FALSE,
       na.action=TRUE)
##########################################################################
#**Effect size for Estimated factor means(package-"effsize")*# gives the
95% confidence interval,
#* for each group latent estimate. Criteria: small effect- d=0.2; medium
effect-d=0.5; Large-d=0.8.
cohen.d(AKADlatentmeans,ENGIlatentmeans)#negligible
cohen.d(AKADlatentmeans,GRUNlatentmeans)#negligible
cohen.d(AKADlatentmeans,SYKPlatentmeans)#small
cohen.d(ENGIlatentmeans,GRUNlatentmeans)#negligible
cohen.d(ENGIlatentmeans,SYKPlatentmeans)#negligible
cohen.d(GRUNlatentmeans,SYKPlatentmeans)#negligible

##########################################################################
#**Sensitivity Analysis - without outliers* #No difference in conclusion
##########################################################################
table(df.without_outliers$Study_Program)
# Analysis Data without outliers
no.outliers<-df.without_outliers[c(1,5:11)]
#Overall fit
cfa.ModelsensOverall.fit <- cfa(cfa.ModelC, data = no.outliers, estimator=
"MLR")
summary(cfa.ModelsensOverall.fit, fit.measures = T, standardized =
T,rsquare=T)
fitMeasures(cfa.ModelsensOverall.fit,
c("cfi.scaled","tli.scaled","rmsea.scaled","srmr"))

##########################################################################
#**Configural*
fit.sensConfigural <- cfa(cfa.ModelC, data = no.outliers, estimator =
"MLR", group = "Study_Program",
                          meanstructure = TRUE)
summary(fit.sensConfigural, standardized = TRUE, fit.measures = TRUE)
fitMeasures(fit.sensConfigural, c("cfi.scaled","rmsea.scaled","srmr"))

##########################################################################
#**Metric*
fit.sensMetric <- cfa(cfa.ModelC, data = no.outliers, estimator = "MLR",
group = "Study_Program",
                      meanstructure = TRUE, group.equal = c("loadings"))
summary(fit.sensMetric, standardized = TRUE, fit.measures = TRUE)
fitMeasures(fit.sensMetric, c("cfi.scaled","rmsea.scaled","srmr"))

#**Compare the model fit indices**
overallfitSens1<-compareFit(fit.sensMetric, fit.sensConfigural)
summary(overallfitSens1,fit.measures =
c("cfi.scaled","rmsea.scaled","srmr"))
##########################################################################
#**Scalar*
fit.sensScalar <- cfa(cfa.ModelC, data = no.outliers, estimator = "MLR",
group = "Study_Program",
```

```
                      meanstructure = TRUE, group.equal = c("loadings",
"intercepts"))
summary(fit.sensScalar, standardized = TRUE, fit.measures = TRUE)
fitMeasures(fit.sensScalar, c("cfi.scaled","rmsea.scaled","srmr"))

#**Compare the model fit indices**
overallfit.sens2<-compareFit(fit.sensScalar, fit.sensMetric)
summary(overallfit.sens2,fit.measures =
c("cfi.scaled","rmsea.scaled","srmr"))
##############################################################################
```

Appendix III: Supplemental Material

A: The Learning Outcome Scale

How satisfied are you with your learning outcomes so far, concerning:

Not satisfied (1) (2) (3) (4) (5) Very satisfied

1. Theoretical knowledge
2. Knowledge of scientific work methods and research
3. Experience with research and development work
4. Discipline- or profession-specific skills
5. Critical thinking and reflection
6. Cooperative skills
7. Oral communication skills
8. Written communication skills
9. Innovative thinking
10. Ability to work independently

B: Correlation Matrix and other Descriptive Statistics

Table 1a

Correlations and other Descriptive Statistics *(n=7488; alpha α= .88)*

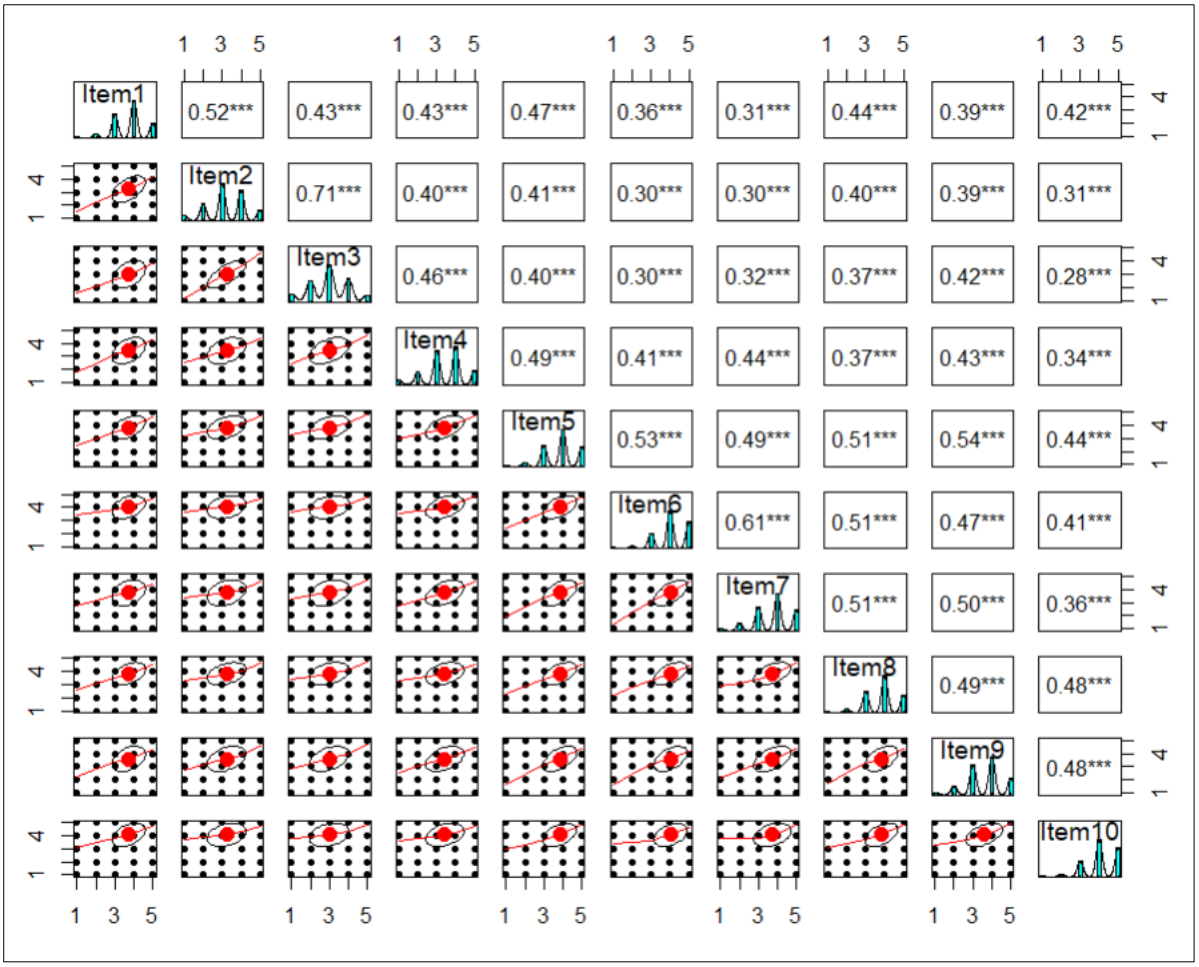
Variable	1	2	3	4	5	6	7	8	9	10
1.Theoretical knowledge										
2. Knowledge of scientific work methods and research	.53									
3. Experience with research and development work	.43	.71								
4. Discipline- or profession-discipline specific skills	.44	.40	.46							
5. Critical thinking and reflection	.47	.41	.40	.48						
6. Cooperative skills	.36	.30	.30	.41	.53					
7. Oral communication skills	.31	.30	.32	.44	.50	.61				
8. Written communication skills	.45	.40	.37	.38	.51	.51	.52			
9. Innovative thinking	.39	.39	.42	.43	.54	.47	.51	.49		
10. Ability to work independently	.42	.31	.28	.35	.44	.41	.37	.48	.48	
Mean	3.73	3.23	2.99	3.43	3.83	4.02	3.73	3.82	3.59	4.05
Standard Deviation	0.88	1.02	1.06	1.01	0.90	0.88	0.99	0.87	0.97	0.89
Skewness	-0.56	-0.22	-0.05	-0.40	-0.62	-0.87	-0.62	-0.58	-0.45	-0.90
Kurtosis	0.41	-0.37	-0.52	-0.21	0.29	0.81	0.06	0.39	-0.06	0.80

Note. Items 1 to 3 are hypothesized to belong to the ***'knowledge achievement'*** LO conceptual sub-dimension, whereas item4 to 10 are hypothesized to belong to the ***'skills achievement'*** sub-dimension.

C: Characteristics of the data set

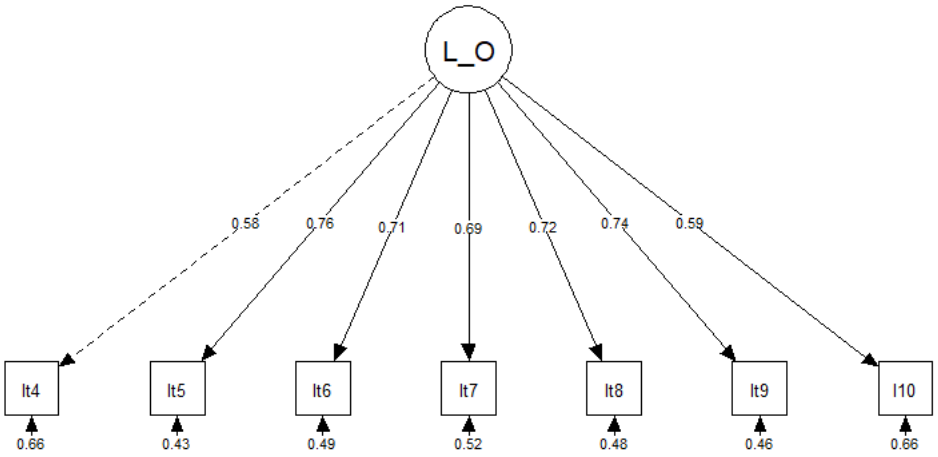
Figure 2

Scatter plot matrix

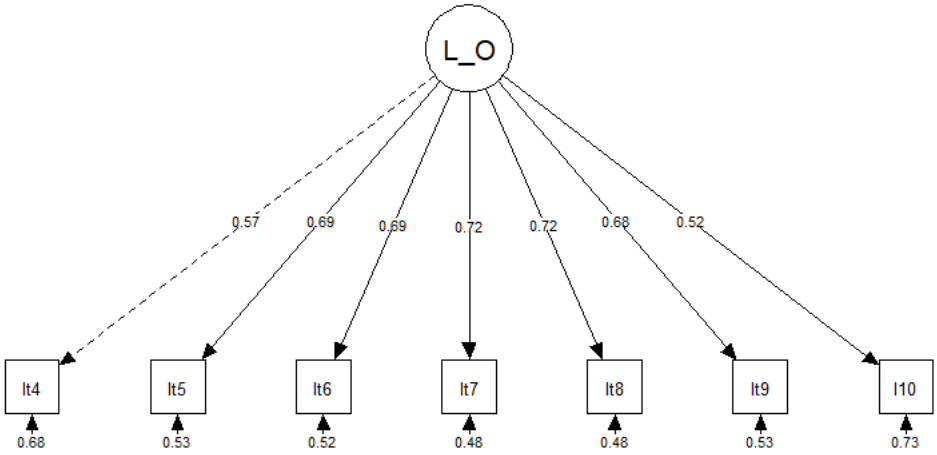


E: Path diagrams for Single group CFA

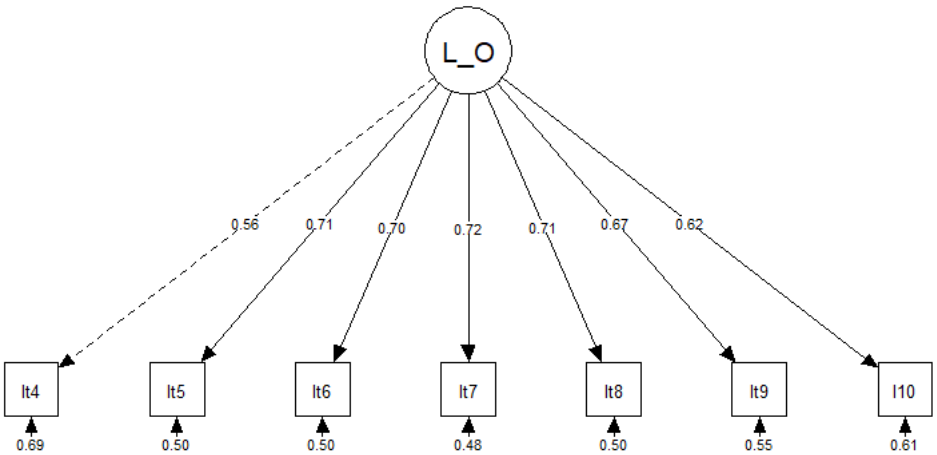
Business Administration Group



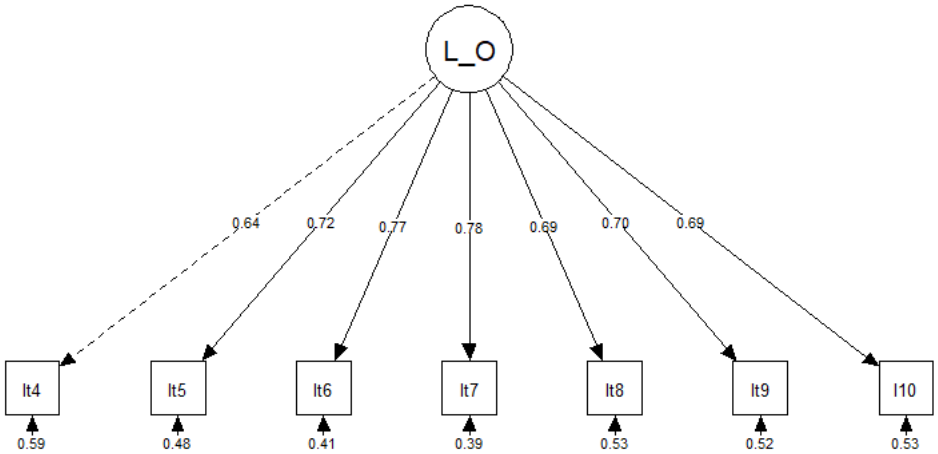
Engineering Group



Teacher Education Group



Nursing Group



G: Pairwise t statistics and Effect Size (Cohen d)

Table 6

Pairwise t statistics and Effect Size (Cohen d) n=7488

	M	B&A	Engineering	Teacher Education	Nursing
B&A	.080		t=4.57 (df=14974, P-value=<.001) d=.075	t=12.20(df=1497, P-value=<.001) d=.199	t=13.60 (df=14974, P-value=<.001) **d=.222**
Engineering	.003			t=7.63 (df=14974, P-value=<.001) d=.125	t=9.03 (df=14974, P-value=<.001) d=.148
Teacher Education	-.126				t=1.40 (df=14974, P-value=.160) d=.0023
Nursing	-.150				

Note. B&A= Business & administration, M= estimated mean, d=Cohen d statistic, p value=0.05, Cohen d statistic criteria= small effect- d=0.2; medium effect-d=0.5; Large-d=0.8.