Aus dem Institut für Didaktik und Ausbildungsforschung in der Medizin
Institut der Ludwig-Maximilians-Universität München

Leitung: Prof. Dr. med. Martin R. Fischer, MME (Bern)

The role of Statistical Literacy for Scientific Reasoning & Argumentation in Medicine

Dissertation

zum Erwerb des Doktorgrades der Medizin

an der Medizinischen Fakultät der

Ludwig-Maximilians-Universität zu München

vorgelegt von

Felicitas Maria Schmidt, MPH

aus
München

2021

Mit Genehmigung der Medizinischen Fakultät

der Universität München

Berichterstatter:	*Prof. Dr. med. Martin R. Fischer, MME*
Mitberichterstatter:	Prof. Dr. Anne-Laure Boulesteix
	Prof. Dr. Jörg Schelling
Mitbetreuung durch den promovierten Mitarbeiter:	*Dr. phil. Markus Berndt*
Dekan:	*Prof. Dr. med. dent. Reinhard Hickel*

Tag der mündlichen Prüfung: 17.06.2021

Eidesstattliche Versicherung

Schmidt, Felicitas Maria

(Name, Vorname)

Ich erkläre hiermit an Eides statt,

dass ich die vorliegende Dissertation mit dem Titel *The role of Statistical Literacy for Scientific Reasoning & Argumentation in Medicine* selbstständig verfasst, mich außer der angegebenen keiner weiteren Hilfsmittel bedient und alle Erkenntnisse, die aus dem Schrifttum ganz oder annähernd übernommen sind, als solche kenntlich gemacht und nach ihrer Herkunft unter Bezeichnung der Fundstelle einzeln nachgewiesen habe.

Ich erkläre des Weiteren, dass die hier vorgelegte Dissertation nicht in gleicher oder ähnlicher Form bei einer anderen Stelle zur Erlangung eines akademischen Grades eingereicht wurde.

München, 24.06.2021 Felicitas M. Schmidt

(Ort, Datum) (Unterschrift Doktorandin)

Content

Tables

Figures

Abbreviations

ANOVA	Analysis of variance
ACTA	Assessment of critical thinking ability
DC	Drawing conclusions
DV	Dependent variable
EB	Epistemological beliefs
EE	Evidence evaluation
EL	Evidence level
GP	General practitioner
IV	Independent variable
LV	Likert value
Max.	Maximum
MME	Master of medical education
MOOCS	Massive open online courses
OGSLQ	Obstetrician-gynecologist statistical literacy questionnaire
Pkt	Punkte (English: points)
PPV	Positive predictive value
RQ	Research question
SDDS	Dual search model
SL	Statistical literacy
SNS	Subjective numeracy scale
SR	Statistical reasoning
ST	Statistical thinking
SRA	Scientific reasoning and argumentation
TOSLS	Test of scientific literacy skills

Zusammenfassung

Hintergrund: Statistical Literacy (SL) und wissenschaftliches Denken und Argumentieren (engl. *scientific reasoning and argumentation*, abgekürzt SRA) sind für die ärztliche Praxis von grundlegender Bedeutung. Statistical Literacy bezeichnet die Fähigkeit, statistische Zahlen und Ergebnisse zu verstehen, zu interpretieren und anzuwenden. Sie wird als wesentliche Voraussetzung für eine angemessene Risikoabschätzung und -kommunikation angesehen. Neben medizinischer Sachkenntnis und Erfahrung bildet SRA zusammen mit Statistical Literacy eine wichtige Grundlage für die evidenzbasierte medizinische Praxis. Studien legen nahe, dass bei Medizinstudierenden und ÄrztInnen beide Fähigkeiten unterentwickelt sind. Ziel dieser Dissertation ist die quantitative Analyse dieser Fähigkeiten bei Medizinstudierenden im Vergleich zu Studierenden aus zwei anderen Fächern (Wirtschafts- und Sozialwissenschaften) sowie zu praktizierenden ÄrztInnen. Die Arbeit soll außerdem einen Beitrag zur Frage leisten, wann, wie und wo ÄrztInnen diese Fähigkeiten erlernen, um sie gezielt fördern zu können.

Methodik: Zu diesem Zweck wurde ein neues Messinstrument entwickelt, das Elemente zur Bewertung der Statistical Literacy in der umfänglichen Definition von Watson (1997) und SRA kombiniert und sich auf die beiden Aktivitäten Evidenzbewertung (engl. *evidence evaluation*, abgekürzt EE) und das Ziehen von Schlüssen (engl. *drawing conclusions*, abgekürzt DC) konzentriert. Der erste Teil der Dissertation besteht aus einer quasi-experimentellen Studie mit $N = 212$ Studierenden der LMU München, in der die Unterschiede in SL und SRA zwischen Studierenden in Abhängigkeit von Studienfortschritt (Bachelor, Master bzw. Vorklinik, Klinik) und Studienfach (Medizin, Sozialwissenschaften, Wirtschaftswissenschaften) untersucht werden. Darüber hinaus wird für logisches Denken, Lernbereitschaft und epistemologische Überzeugungen kontrolliert. Der zweite Teil dieser Dissertation umfasst eine quasi-experimentellen Studie mit $N = 71$ deutschsprachigen ÄrztInnen. Für diesen Zweck wurde das Messinstrument um einen umfangreichen demografischen Abschnitt erweitert. Diese Studie geht der Frage nach, wie, wann und wo SL und SRA erlernt wurden. Im dritten Teil dieser Dissertation wurden weitere Auswertungen mit den ÄrztenInnen aus dem zweiten Teil und den Medizinstudierenden aus dem ersten Teil durchgeführt, so dass ein domänenspezifischer, quantitativer Vergleich entstand.

Ergebnisse: Im ersten Teil der Doktorarbeit zeigt sich ein signifikanter Haupteffekt von Studienfach auf Statistical Literacy, wobei Studierende der Sozialwissenschaften weniger Punkte erzielen als Medizin- bzw. Wirtschaftswissenschaftsstudierende. Keine Unterschiede zeigten sich bezüglich SRA.

Im zweiten Teil zeigen die ÄrztInnen ein mittleres Niveau in SL und SRA. Die ÄrztInnen geben an, ihre Fähigkeiten hauptsächlich autodidaktisch entwickelt zu haben. Die aktive Beteiligung an der Forschung scheint bei der Entwicklung der Fähigkeiten entscheidend zu sein: Die Anzahl der Veröffentlichungen und die für die Forschung aufgewendete Zeit sagen SL weitgehend voraus. SRA-Fähigkeiten sind stark mit der Art der medizinischen Doktorarbeit assoziiert (experimentell, klinisch) und mit Arbeitserfahrung in der Forschung.
Kein Unterschied zeigt sich zwischen Medizinstudierenden und ÄrztInnen (Teil III) bezüglich Statistical Literacy und der Evidenzbewertung. Jedoch unterscheiden sich die Gruppen signifikant in Bezug auf Schlussfolgern. Medizinstudierende erreichen dabei höhere Scores als ÄrztInnen. Medizinstudierende und ÄrztInnen weisen ähnliche Leistungsmotivation und logisches Denken auf, zeigen aber Unterschiede in einigen epistemologischen Überzeugungen.

Zusammenfassung: Medizinstudierende weisen im Vergleich zu Studierenden der Sozialwissenschaften eine bessere und im Vergleich zu Studierenden der Wirtschaftswissenschaften eine ähnliche Statistical Literacy auf. Die SRA-Fähigkeiten unterschieden sich nicht zwischen den Studienfächern, scheinen aber von unterschiedlichen epistemologischen Überzeugungen geprägt zu sein. Da bei Medizinstudierenden in der klinischen Phase des Studiums ein Rückgang der Statistical Literacy beobachtet wurde, wird in dieser Arbeit ausdrücklich für die Förderung dieser Fähigkeiten, zum Beispiel durch aktive Beteiligung an Forschungsaktivitäten, argumentiert. Dies schien eine wichtige Rolle bei der Entwicklung von SL- und SRA-Fähigkeiten bei ÄrztInnen zu spielen. Diese Dissertation könnte somit zur Rechtfertigung der Umsetzung einer systematischen Förderung während der universitären medizinischen Ausbildung und während der Weiterbildungszeit beitragen.

Summary

Background: Statistical literacy (SL) and scientific reasoning and argumentation (SRA) skills are considered as fundamental for professional practice. Statistical literacy can be defined as the ability to understand, interpret, and use statistical numbers and its results. It is regarded as an essential prerequisite for risk estimation and communication. Together with SRA skills, SL can be seen as an essential basis for evidence-based practice. Several studies suggest that in medical students and physicians both skills are underdeveloped. The aim of this dissertation is to examine these skills in medical students, in comparison to students from two other domains (economics, social sciences) and in practicing physicians. Furthermore, it should contribute to the discussion, how, when, and where physicians acquire these skills in order to foster them.

Methods: For this purpose, a new measurement tool was created combining items assessing SL in the comprehensive definition of Watson (1997) and SRA. It will focus on the two epistemic activities evidence evaluation (EE) and drawing conclusion (DC). Part I constitutes a quasi-experimental 2x3 study with $N = 212$ students from the LMU Munich, investigating the differences in students depending on study progress (undergraduate, graduate) and study domain (medicine, social sciences, economics). Additionally, it was controlled for logical reasoning, willingness to learn, and epistemological beliefs. In the second part, the measurement survey included an extensive demographic section and applied to $N = 71$ German-speaking physicians. In the third part of this dissertation, further analyses were carried out with the physicians from the second part and the medical students from the first part, resulting in a domain-specific, quantitative comparison.

Results: In Part I, a significant main effect of study domain was found, but only regarding SL, with Social Sciences students scoring lower than other students. No differences were found regarding SRA. In Part II, physicians showed a medium level of SL and SRA. Participants indicated to have developed their skills mostly in autodidactic learning activities. The active involvement in research seemed decisive: The number of publications and time spent in research predicted SL to a large extent. SRA skills were associated with the type of MD-thesis (experimental, clinical) and working in research.

Comparing medical students with physicians (Part III), statistical literacy did not differ significantly across the different stages of medical education, as well as in evidence evaluation scores. However, they differed significantly regarding DC. Medical students and physicians did not differ in their achievement motivation or their logical reasoning skills but displayed differences in the some of the epistemological beliefs.

Conclusion: Medical students demonstrate SL skills that were comparable to those of Economics students and superior to those of Social Sciences students. SRA skills did not differ across domains but seemed to be shaped by different epistemological beliefs. Due to an observed decline in statistical literacy among medical students in the clinical phase of their studies, it will be argued explicitly for the promotion of these skills, for example through active participation in research activities. This seemed to play a vital part in the development of SL and SRA skills. This dissertation could thus contribute to the justification of the implementation of systematic fostering during formal medical education and during residency.

1. Statistical Literacy in a Knowledge Society

In the last centuries, driven by strong disruptions, societies and economies gradually transitioned from the agricultural to the industrial age. Agricultural societies were traditionally very labor intensive and skilled workers were mostly trained on the job. The fact that skills promote economic growth, rather than labor and physical capital, was already observed between 1909 and 1949 (Bell, 1976). Due to the development of the internet, the revolution of communication and rise of digital technologies, often described as the Third Industrial Revolution, the requirements for skilled workers in all industries but mostly in the service sector have changed (Välimaa & Hoffman, 2008). In this often-postulated knowledge-society, not only are the driving forces of innovation more and more deducted from research and development, but also a growing proportion of GDP and employment is attributable to the field of knowledge rather than agriculture or mere mass-production (Bell, 1976). Due to this development, the foundation of productivity shifted to the means of knowledge generation and information processing (Castells, 1996). These developments implement new requirements for skilled workers (Välimaa & Hoffman, 2008). To actively take part in a society that collects, analyses and evaluates data and largely bases its decisions on it, specific aptitudes are required. Several studies have been conducted in order to identify the most relevant requirements (i.e. key competencies). For example, the OECD project: DeSeCo 2001, i.e. *Definition and Selection of Competencies* (Rychen & Salganik, 2003). Its goal was to construct an overarching conceptual framework required for the development of essential abilities in a lifelong learning setting. These skills are not limited to cognitive skills and knowledge, but will also take the importance of attitude, motivations and values, into account (Rieckmann, 2012; Rychen & Salganik, 2003).

In medicine, a field in which new evidence is generated at a high rate, mere knowledge is not sufficient for physicians to ensure evidence-based and individualized patient care throughout their professional lives. Thus, teaching medicine cannot focus exclusively on the memorization of diagnostic and treatment algorithms. It has to promote the development of key qualifications for their evidence-based practice and participation in a science-based society. Among others, statistical literacy (SL) and scientific reasoning and argumentation (SRA) skills are considered as prerequisites (Callingham & Watson, 2017; Gigerenzer & Gaissmaier, 2008).

Statistical literacy can be described as the ability to understand statistical information and to apply it in the decision-making process. It is indispensable in many professional contexts, like

medicine, and in a society increasingly based on quantitative knowledge and evidence (Callingham & Watson, 2017). Statistical literacy can be defined as the ability to critically reflect on statistics as evidence in arguments (Shield, 1999). As such, it is connected to SRA skills (Fischer et al., 2014) which constitute the basis for evidence-based decision-making (Sedlmeier & Gigerenzer, 2001). To prepare university students for their academic and professional life, the development of SL and SRA is an aim of the curriculum in many science-based study domains, including medicine, as well as economics and other social sciences. Physicians, for instance, must be informed about the risks and benefits different treatments offer for their patients (Gaissmaier & Gigerenzer, 2008). It can also be considered as a prerequisite for effective risk communication (Nelson, Reyna, & Fagerlin, 2008; Reyna, Nelson, Han, & Dieckmann, 2009).

Unfortunately, a collective statistical illiteracy has been observed among physicians (Lipkus, Samsa, & Rimer, 2001). Likewise, SRA skills needed for evidence-based practice are underdeveloped (Sedlmeier & Gigerenzer, 2001; Anderson, Gigerenzer, Parker, & Schulkin, 2014).

Although several instruments assessing SL and SRA have been described in the literature, none of them combine the measurement of both skills across study domains. Furthermore, attitudes and motivations, as indicated by Rychen and Salganik (2003), will be taken into account in the scope of this dissertation.

Thus, this dissertation aimed at gaining insights into in the current development of SL and SRA in medical students and physicians. Furthermore, it should contribute to the discussion, how, when and where physicians acquire these skills and provide insights in the relationship of SL and SRA in order to foster these skills in the medical domain.

For this purpose, this dissertation is structured as follows: After the introduction, chapter 2 (2. Theoretical Background) creates the theoretical framework providing an overview of the underlying concepts and definitions, which will form the basis for the further parts.

The first part explores SL and SRA in $N = 212$ students, comparing medical students to students from the domains of economics and social sciences. Additionally, it controlled for logical reasoning, willingness to learn, and epistemological beliefs. This should contribute to the question in what manner SL and SRA skills are intertwined and highlight the role of these factors for these skills.

In the second part, SL and SRA are assessed in physicians. This part focuses not only on skill assessment but adds due to its demographic section also information on skill development

during the professional career. The third part compares medical students with physicians allowing for an intra-domain comparison and model building. Fragments of Part I and Part II are currently in the publishing process in peer reviewed journals.

Based on these findings, the last chapter (6. Discussion) then attempts to answer and discuss the research questions introduced at the end of chapter 2.

The present dissertation closes with a supposition and an outlook on the further development of SL and SRA in the medical domain.

2. Theoretical Background

2.1. Introduction

This chapter introduces the two skills SL and SRA, provides insights in their development and gives an overview of the various measurement tools in the existing literature. Furthermore, it will summarize the prevailing concepts of a theoretical model of scientific reasoning. It will be contested to what extent SL and SRA are intertwined. This part concludes with a detailed list of research questions and hypotheses.

2.2. Statistical Literacy

2.2.1. Concepts and Definitions

Everyone who took a statistics class in higher education or faced a statistical problem in everyday life, is aware that mere knowledge of statistical concepts is just half the battle to solving the problem. Statistical knowledge (SK) can be described as the understanding of statistical models and notions, such as probability, uncertainty, distribution, sampling and association and provides the basis for statistical problem solving (Broers, 2006). The other half of the battle constitutes in a comprehensive notion, whose definition evolved over time: Statistical literacy (SL).

One of the early definitions described SL as scientists' ability to use quantitative language (Walker, 1951). Later, it has been defined as the ability to comprehend and analytically assess statistical numbers and results in everyday life and to understand and acknowledge statistical input in private and professional decision-making (Wallman, 1993). It is based on numeracy which refers to the ability of dealing with mathematical operations (Peters, 2012).

Statistical literacy involves statistical knowledge about relative risks and conditional probabilities (Covey, 2007), the understanding of multivariate connections (Gal, 2002), correlations (McKenzie, 2004) and confounding biases (Shield, 1999).

Important prerequisites for SL are familiarity with statistical concepts, vocabulary, and symbols and the concept of probability as a measure of uncertainty.

Watson (1997) categorized SL in a hierarchical manner with basic numeracy and understanding of probabilities as the lowest level. The intermediate level constitutes the contextual comprehension of statistical language. The highest level is represented by a critical attitude towards statistical arguments (Watson, 1997).

Watson and Callingham (2003) empirically found a hierarchical model of SL with six different levels: (1) *idiosyncratic*: personal beliefs prevail, (2) *informal*: intuitive, non-statistical

engagement, (3) *inconsistent*: with a rather qualitative involvement that offer not necessarily the correct justifications for a right conclusion, (4) *consistent, non-critical*: sufficient understanding of the problem and the context, however, not being questioned, (5) *sophisticated*: in well-known contexts critical thinking and suitable use of vocabulary, (6) *critical mathematical*: extensive understanding and application of statistical formulas in relation to context and critical reasoning. Statistical literacy based on this model can be best described as a thick rope consisting of two knitted threads: mathematical comprehension of the content and involvement in the (social) context the problem is based in (Tognolini, 1996).

Based on Watson and Callingham (2003), Ben-Zvi and Garfield (2004) classified SL in three levels and relating it to the notions statistical thinking and statistical reasoning.

Statistical reasoning is equally grounded on statistical knowledge. However, statistical reasoning can be seen as the way a person thinks about statistical information (Garfield, 2002). Statistical thinking was described by Snee (1990) as the "thought process, which recognises that variation is all around us and present in everything we do, all work is a series of interconnected processes, and identifying, characterizing, quantifying, controlling and reduction variation provide opportunities for improvement." (p. 116). In the model by Ben-Zvi and Garfield (2004), the lowest level of mental involvement with statistics includes only fundamental aptitudes for the conventional employment of statistics and its results. At a more advanced level, statistical reasoning and statistical thinking were determined. The first is described as the understanding of statistical procedures, and the latter as the critical evaluation of statistical procedures and problems (Ben-Zvi & Garfield, 2004). Analogously, Gal (2004) also embedded the concepts of statistical thinking (i.e. critique and communicate statistical results) and statistical reasoning (i.e. comprehension of statistical results) in his definition of SL.

In contrast to the abovementioned definitions, there is also research that contests this hierarchical order with SL overstretching statistical reasoning and statistical thinking: delMas (2002) perceived SL in the sense of understanding results, graphs and concepts. As such, SL constitutes a prerequisite for statistical reasoning, which engages in the process and the context in which these results were created (delMas, Ooms, Garfield, & Chance, 2006). In the Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report (2016), the definition of the term SL was superseded by the notion statistical thinking (Franklin et al., 2005).

The definition of SL used in this dissertation follows Watson (1997) comprising not only the understanding of probabilities and basic concepts, but also the contextual comprehension of

statistical language. Moreover, it requires a critical stance towards statistical information. The definition by Ben-Zvi and Garfield (2014) can be applied as well when taking reasoning and thinking into account.

2.2.2. Development of Statistical Literacy

As already outlined, the complex nature of SL and its interconnected components reasoning and thinking, has created a challenge in fostering and developing these skills in school and higher education. Statistical literacy, statistical reasoning and statistical thinking are all grounded on statistical knowledge, which is, can and must be taught and fostered in formal education. However, following Broers (2006), to become a statistically literate person, it requires a critical attitude, consideration and intelligence. As such, teaching statistics seems to be merely sufficient.

Additionally, teaching statistics is not a trivial task. The teaching method is regarded as a relevant factor. Often, the focus of statistical education lies on mathematical and procedural aspects of knowledge (e.g. performance of calculations). This could impede students from developing problem-solving skills and eventually SL in their respective fields (Garfield, 1995).

From the students' perspective, statistics is considered a difficult subject (Zieffler et al., 2008). Cognitive and demographic factors – such as gender, prior knowledge, mathematical skills and attitudes towards statistics – were found to predetermine influencing the students' performance in statistic courses. Particularly, the attitudes towards statistics were observed to differ across the subjects with graduate economics students being most positive and valuing statistics for their future career (Griffith, Adams, Gu, Hart, & Nichols-Whitehead, 2012).

Thus, initiatives and intervention studies have sought to improve the learning experience of statistics in schools (Budgett & Rose, 2017), higher education (Lloyd & Robertson, 2012) and through a variety of organizational efforts (Franklin et al., 2005; Shield, 2017).

2.2.3. Measurement of Statistical Literacy

Reflecting the subtle dissimilarities in these definitions, there are a myriad of tests and measurement tools assessing SL. An attempt of categorizing the testing methods is the differentiation between subjective and objective measurements. Subjective measures allow for an unfiltered self-estimation, take less time and are better supported by participants (Dolan, Cherkasky, Li, Chin, & Veazie, 2016). Nonetheless, research suggests that the self-assessment of one's own SL is often inaccurate, which may negatively influence the handling of statistical data or lead to errors in decision-making processes (Fagerlin et al., 2007). Physicians, for

example, tend to overestimate their skills, because it is socially desirable for them to be well informed about risks and probabilities in order to weigh treatment alternatives (Anderson et al., 2014). Objective measurement tools are, *inter alia*, ability tests. They are often associated with negative experiences and the unwillingness to participate and consequently higher attrition rates. Furthermore, when completing test instruments online, it cannot be controlled for the use of calculators (Fagerlin et al., 2007). Subjective and objective test results have been found to correlate well in Fagerlin et al. (2007) comparing the Subjective Numeracy Scale (SNS) with the Expanded Numeracy Scale by Lipkus et al. (2001). However, these findings were not reproduceable (Hess, Visschers, Siegrist, & Keller, 2011; Rolison, Wood, Hanoch, & Liu 2013), suggesting that subjective and objective measurement tools assess different concepts.
In this dissertation, subjective and objective measurement scales and their correlation were calculated for students and physicians.

Objective measurement tools often only assess a certain trait, usually considering one of the three levels of Watson (1997) or the specific domain, e.g. medical context (Anderson et al., 2014). A prominent example is the Three Item Test (Schwartz, Woloshin, Black, & Welch, 1997). It contains a conversion from a percentage to a frequency and vice versa and an estimation of basic probability.
Lipkus et al. (2001) expanded this test with the conversion of probabilities to proportions. These concise tests can be considered less suitable for highly educated samples due to ceiling effects. More than half of the participants in such samples received full scores (Hanoch, Miron-Shatz, Cole, Himmelstein, & Federman, 2010). A further expansion was created by Cokely et al. (2012) with the Berlin Numeracy Test (4 Items; Cronbach's $\alpha = 0.59$). It was validated with different samples from various cultural contexts across 15 countries and proved to be suitable for higher educated samples (Cokely, Galesic, Schulz, Ghazal, & Garcia-Retamero, 2012).

A prominent example of statistical ability tests in the medical domain is the Medical Data Interpretation Test (Schwartz et al., 2005, 6 Items, Cronbach's $\alpha = .71$). Participating physicians scored higher overall than others with postgraduate degrees (89 out of 100 score points). More specific for a medical discipline is the Obstetrician-Gynecologist Statistical Literacy Questionnaire (OGSLQ, Anderson et al., 2014, 14 Items; Cronbach's $\alpha = .53$). The OGSLQ, taken by 200 US-American obstetricians and gynecologists, comprised 14 questions in the respective context in the three different fields: numerical facts, statistical concepts and questions on numerical relationships. Physicians scored well on the latter, with 90% being able to convert frequencies to probabilities, lower on statistical concepts, with 49% calculating

correctly the positive predictive value and worse on numerical facts with 19% estimating correctly the number of women in the US with cancer (Anderson et al., 2014). Similar results were found by Gigerenzer and Wegwarth (2008), showing that 79% being unable to interpret the positive predictive value.

Gigerenzer et al. (2008) demonstrated that many physicians struggle with the statistical implications of mammography screenings leading to an overdiagnosis bias, i.e. the unnecessary treatment of potentially suspicious findings. Moreover, they found that 50% of the participants thought that false positive test results in HIV testing do not exist, only two out of 20 urologists had sufficient knowledge about the reliability of a PSA-test, and 85% of advanced medical students drew wrong conclusions from the positive predictive value of four different early detection methods (Gigerenzer, Gaissmaier, Kurz-Milcke, Schwartz, & Woloshin, 2008).

Furthermore, physicians have been observed to favor treatments when evidence was presented in relative risks instead of absolute risks, suggesting an inability to draw the right conclusions from certain forms of risk representation (Covey, 2007).

Generally, there is more research focusing on the SL of physicians than on residents or medical students. A study in Greece showed that only two out of 153 medical residents answered all seven knowledge questions of a self-designed test correctly, whereas almost 20% answered all questions incorrectly (Msaouel et al., 2014). Another study assessed the numeracy of medical students and residents and found students with poor numeracy being more likely to misjudge risks of different treatment alternatives with increasing confidence in treatment recommendations over the duration of medical school (Johnson et al., 2014). Residents' average scores in a biostatistics test differed significantly from that of fellows (41.4% vs. 71.5%), although 95% found the questioned concepts relevant for the comprehension of scientific literature (Windish, Huot, & Green, 2007).

Overall, physicians' SL is not ideal (Anderson et al., 2014), however, it can be seen as comparable to that of other educated samples (Lipkus et al., 2001; Okamoto et al., 2012) and was found superior to that of residents in research training (Windish et al., 2007) or medical students (Johnson et al., 2014).

2.3. Scientific Reasoning and Argumentation

2.3.1. Notion and Definition

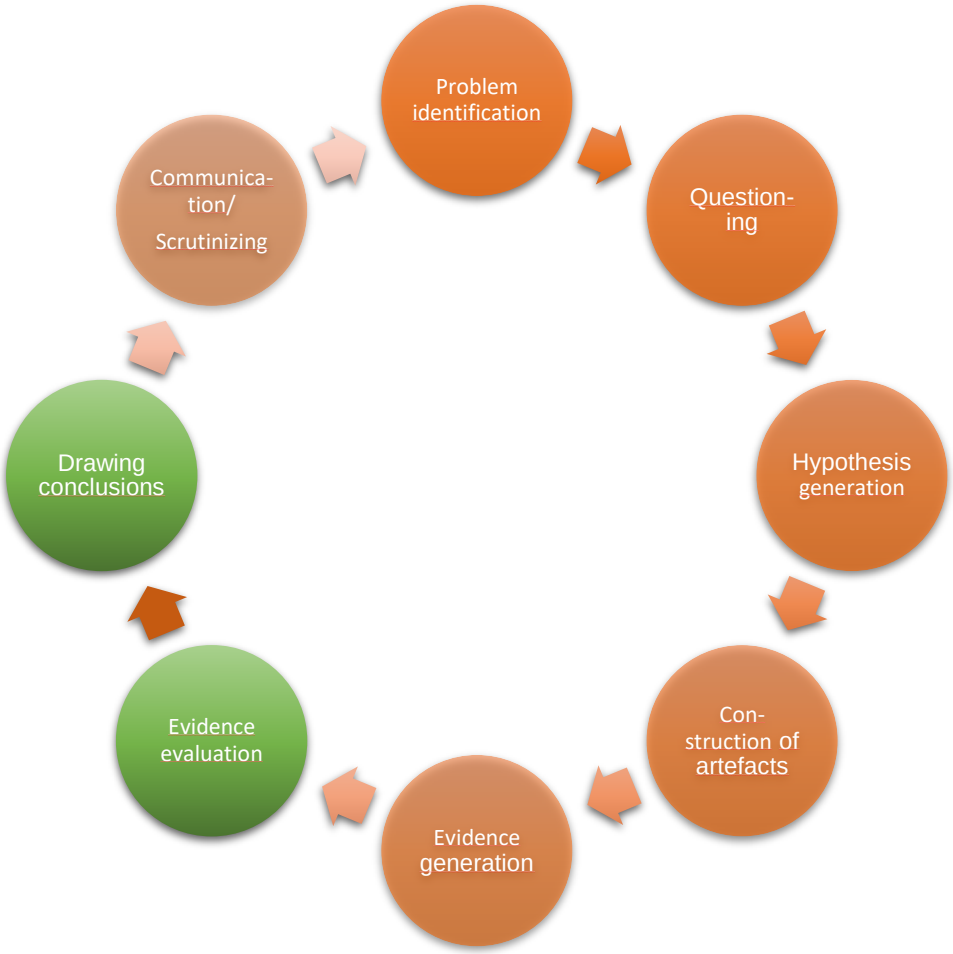
Analogously to SL, scientific reasoning and argumentation (SRA) skills can be considered as essential abilities for participating in a knowledge-based society and thus regarded as indispensable targets of science education (Fischer et al., 2014; Osborne, 2010). Like SL, SRA skills have become increasingly important for informed decision-making, not only in a scientific context, but also in everyday life. In this dissertation, the notion *skill* is used in a very general sense but understood as being different to intelligence (since it can be trained) and conceptual knowledge, e.g. statistical knowledge.

Engelmann et al. (2016) hypothesized scientific reasoning as the composition of three principal elements. First, they considered scientific reasoning as a course of scientific discovery (Fischer et al., 2014). Secondly, they understood scientific reasoning concentrating on argumentation (Osborne, 2010) and last, it could be regarded as the attempt to understanding the nature of science, as described by Lederman (2007).

Following the framework proposed by Fischer et al. (2014), scientific reasoning and argumentation (SRA) can be defined as a set of competencies, comprehending and applying scientific working methods and their results when solving problems (Hetmanek, Engelmann, Opitz, & Fischer, 2018; Rudolph & Horibe, 2016). It "includes the knowledge and skills involved in different epistemological activities (…) in the context of three different modes (advancing theory building about natural and social phenomena, science-based reasoning in practice, and artifact-centered scientific reasoning)" (Fischer et al., 2014, p. 39). These epistemological activities are: (1) Problem identification, (2) Questioning, (3) Hypothesis generation, (4) Construction of artefacts, (5) Evidence generation, (6) Evidence evaluation, (7) Drawing conclusions and (8) Communication and Scrutinizing (Figure 1). It can be either displayed as iterative circle, as in Figure 1, or as individual epistemological activities that do not necessarily have to be conducted in a specific order. This dissertation focuses on evidence evaluation (EE) and drawing conclusions (DC).

Figure 1

Framework of Scientific Reasoning and Argumentation (SRA)



Note: Adapted from Fischer et al. (2014)

2.3.2. Development of SRA

Scientific reasoning and argumentation, which Inhelder and Piaget (1958) assumed to constitute the highest form of human thinking, is considered to develop already in childhood. Children understand science as actions and their consequences, representing Level 1, while adults more often understand science as offering explanations by hypothesis testing (Level 2). However, few people reach Level 3, understanding science as a repetitive itinerary of theory, testing, and revision (Carey & Smith, 1993).

The so-called dual search model (SDDS), developed by Klahr and Dunbar (1988), defined hypothesis generation, evidence generation and evidence evaluation as iterative circle of scientific reasoning. Other studies found that SRA did not differ fundamentally between adulthood and childhood, since both tended to be context dependent (Koslowski, 1996; Zimmerman, 2000) and domain-specific knowledge and domain-general skills seemed to

interact (Opitz, Heene, & Fischer, 2017). However, this amalgamate was found to be more present in adult SRA, which was observed to be more domain-specific (Kruglanski & Gigerenzer, 2011) and influenced by prior knowledge and theoretical biases (Dunbar, 1995).

2.3.3. Measuring SRA

The assessment of SRA skills is highly dependent on the theoretical framework the test instrument is embedded in. For scientific reasoning, in the sense of scientific discovery, there are currently more test scales available than as regarded in the sense of nature of science or argumentation (Opitz et al., 2017).

Building on the framework established by Fischer et al. (2014), coding methods for the quantitative and qualitative evaluation of epistemological activities were developed and effectively employed in several studies and domains, e.g. teacher education (Csanadi, Eagan, Kollar, Shaffer, & Fischer, 2018), social work (Ghanem, Kollar, Fischer, Lawson, & Pankofer, 2016) and medical education (Lenzer, Ghanem, Weidenbusch, Fischer, & Zottmann, 2017).

Since epistemological modes and activities were considered relevant across disciplines, Hetmanek et al. (2018) introduced the conception of cross-domain SRA skills, i.e. skills applicable in several domains, to supersede the dichotomy of domain-general and domain-specific SRA skills. In this dissertation, the notion *domain* denotes a field of study.

Furthermore, in the last 20 years, the conceptualizations of SRA have shifted to multidimensional constructs, which is reflected by a multitude of assessment methods, with items displaying the whole range from closed (multiple choice questions, e.g. (Gormally, Brickman, & Lutz, 2012)) to open (interviews, free text production (Timmerman, Strickland, Johnson, & Payne, 2011)).

In a review by Opitz et al. (2017), 18 out of 38 examined test instruments contained evidence evaluation (EE) and drawing conclusions (DC) differing in their target group (from elementary school children to university students), their item design and the domain of the respective context. For example, the Assessment of Critical Thinking Ability (ACTA) survey used a mix-method approach to assess EE and DC in university students (White et al., 2011), while the Test of Scientific Literacy Skills (TOSLS) measures evidence generation, EE, DC and communicating and scrutinizing among other skills in a cross-domain setting in university students in a multiple choice format (Gormally et al., 2012).

Little research has been conducted so far on the comparative assessment of SRA skills of higher education students of different domains and study phases. A study showed little variation in

SRA skills in students of different study locations, domains and study phases (Lin, Wei, & Mollohan, 2016).

Although designed in medical context, White et al. (2011) tested the ACTA in biology and chemistry students. They assessed three essential critical thinking skills: the ability to integrate conflicting studies, to plan and propose studies to solve these arising conflicts and to assume further interpretations of these studies. These capabilities were found to improve over the course of studies (White et al., 2011).

In Germany, the urgent need to foster SRA skills in medical students has been explicitly expressed (Wissenschaftsrat, 2014). Similarly, a collective statistical illiteracy has been observed (Gaissmaier & Gigerenzer, 2008). With that, the requirement for standardized assessment of these skills in medical students in different stages of their medical education and in comparison to other domains with methodological education and physicians arose.

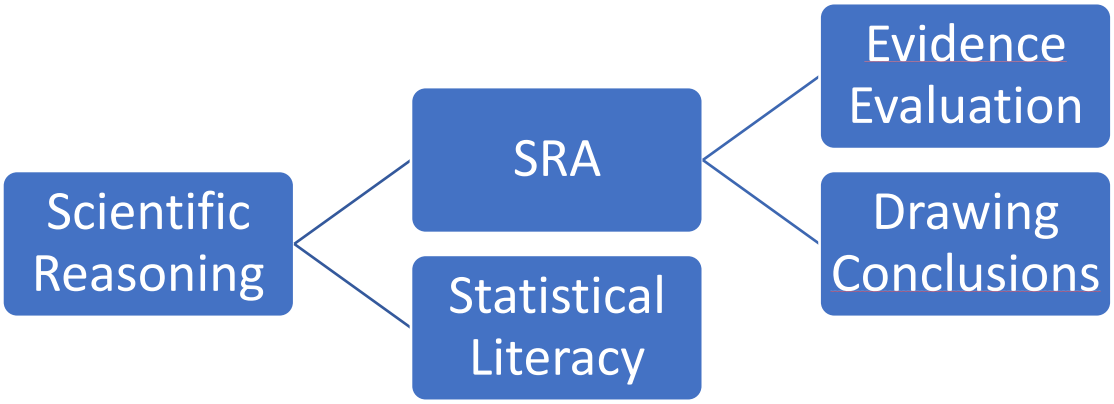
2.4. Theoretical Framework of Scientific Reasoning

Various studies suggest an intertwining of SL and SRA in the overarching construct of scientific reasoning. Several studies argued that SL is needed to evaluate scientific evidence (Anderson, Williams, & Schulkin, 2013; Shavelson & Huang, 2003; Watson & Callingham, 2003). The definition of SL proposed by Anderson et al. (2014) also advocated for SL to be considered as a prerequisite for scientific reasoning. As such, it was described as the use and interpretation of statistical numbers in the context of science. Franklin et al. (2005), on the other hand, hypothesized that SL itself could encompass SRA skills.

In this dissertation, SL and SRA are preliminary hypothesized as two separate constructs, both required in the process of scientific reasoning (Figure 2).

Figure 2

Contextual Framework

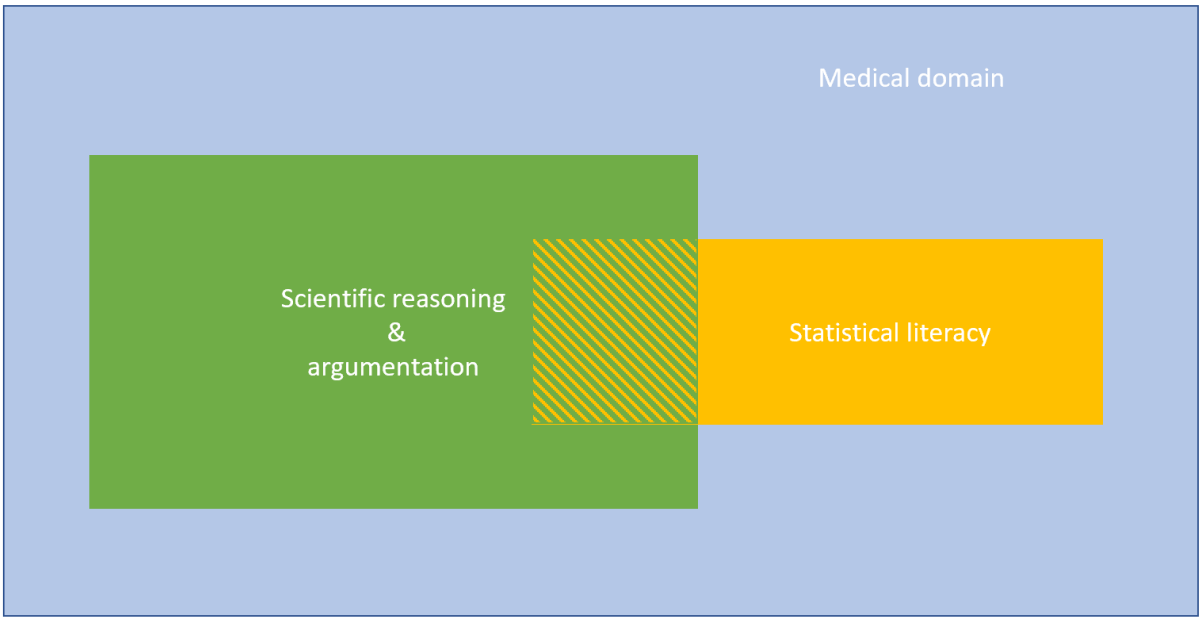


Students' performance in SL and SRA is thought to be influenced by a variety of factors. From studies assessing numeracy, it is known that intelligence can be considered as a strong predictor of SL (Cokely et al., 2012; Frederick, 2005; Lipkus et al., 2001).

Not only the ability to reason, but also the individuals' epistemological beliefs may impact the development of understanding of e.g. statistical concepts or the ability to evaluate evidences and draw conclusions (Diamond & Stylianides, 2017; Ernest, 2011; Garfield, 1995; Mius, 2004). Epistemological beliefs are ideas about knowledge and knowing, which operate mostly during the process of learning new information and skills (Hofer & Pintrich, 2012). Epistemological beliefs may also have an effect on the understanding of the context, the evidence is presented in, and their evaluation in very complex contexts (Porsch & Bromme, 2011) and be linked to achievement motivation (Urhahne, 2006) and willingness to learn.
These factors may help to shed light on the relationship between SL and SRA in the sense that they could potentially influence both in different ways (overlap in Figure 3).

Figure 3

Preliminary Model



2.5. Research Questions & Hypotheses

This dissertation aims at analyzing SL and its relevance for and its relationship with SRA in the medical context. It aspires to contribute to the better understanding of these skills and how to foster them in the medical context. Guiding research questions of this dissertation are:

1. **Statistical Literacy and SRA in students**
   - 1.1. To what extent is statistical literacy developed in medical students in comparison to that of students of economics and social science?
     - 1.1.1. Hypothesis 1a: Economics and social sciences students show higher SL than medical students due to their more extensive science curriculum.
     - 1.1.2. Hypothesis 1b: Graduate students show higher SL than undergraduate students due to an increase in their skills during university studies.
   - 1.2. To what extent is SRA developed in medical students in comparison to that of economics and social science students?
     - 1.2.1. Hypothesis 2a: Economics and social sciences students show higher SRA skills than medical students due to their more extensive science curriculum.
     - 1.2.2. Hypothesis 2b: Graduate students show higher SRA skills than undergraduate students due to an increase in their skills during university studies.

2. **Statistical Literacy and SRA in physicians**
   2.1. To what extent is statistical literacy and SRA developed in physicians?
   2.2. Are SL and SRA skills of physicians' superior to those of medical students?
      2.2.1. Hypothesis 3a Physicians' SL is superior to that of medical students.
      2.2.2. Hypothesis 3b Physicians' SRA is superior to that of medical students.
3. **Education and Development**
   3.1.How, where, and when do physicians develop SL and SRA skills?
   3.2.Which demographic factors are related to the development of SL and SRA?
4. **Model Building**
   4.1. In what manner are statistical literacy and scientific reasoning and argumentation skills intertwined?
   4.2. What is the role of other influencing factors such as logical reasoning, EB, and willingness to learn for SL and SRA skills? (explorative)

To answer these questions, a unique measurement tool was created combining items that assess SL in the comprehensive definition of Watson (1997) and SRA, focusing on the two epistemological activities evidence evaluation and drawing conclusion. It constitutes a quantitative analysis of these skills in two different samples: medical students, in comparison to economics and social sciences students and physicians.

For this purpose, this dissertation is organized in three successive parts. Part I compares the student sample, answering research questions 1 and 4. In the second part, the measurement tool is applied to a physician sample targeting research questions 2 and 3. A third part comprises the analysis of these skills solely in the medical domain, comparing medical students and physicians. All three sections contribute to research question 4., shedding light on the relationship of SL and SRA.

3. SL & SRA across Study Domains – Part I

3.1. Introduction

The first part of this thesis aims at investigating research question 1, namely how SL and SRA are developed in students. It analyses these skills across different domains with varying degrees of statistical and methodological content (Medicine, Social Sciences, Economics) and depending on study progress (undergraduate, graduate). Furthermore, research question 4.2. looking at the role of influencing factors such as logical reasoning, epistemological beliefs and willingness to learn for SL and SRA skills is addressed. Research questions will be discussed in section 6. Discussion, after all results have been presented.

3.2. Methods

3.2.1. Design and Sample

For this section, a 3x2 quasi-experimental, cross-sectional design was applied to the study domain (Medicine, Social Sciences, Economics) and study phase (undergraduate, graduate) as independent variables. From the LMU University, N = 212 students were included. 76 were affiliated with social sciences (including educational sciences, sociology, and psychology), 87 with medicine, and 49 with economics and management sciences (Table 1). These domains have varying degrees of statistical and methodological content in their curricula at the LMU Munich.

Table 1

Demographic Description of the Student Sample

	study phase	*N*	Gender		Mean Age
			Female	Male	(±*SD*)
Medicine	undergraduate	44	25	19	23.25 ± 3.59
	(Vorklinik)		24	19	
	graduate (Klinik)	43			
Social Science	undergraduate	26	25	1	23.89 ± 3.07
	(Bachelor)		44	6	
	graduate (Master)	50			
Economics	undergraduate	33	18	15	22.41 ± 2.54
	(Bachlor)		10	6	
	graduate (Master)	16			
Total		212	146	66	23.31 ± 3.23

Demographic parameters were assessed with special focus on gender, age, and study program and progress. For medical students, the progress of their potential doctoral studies, and whether they had already been using statistical techniques, were assessed.

3.2.2. Instrument

An extensive test instrument was designed uniquely combing the assessment of SL, SRA skills, demographics and control variables. For the measurement of SL, multiple choice items from different, already validated tests were used. SRA was assessed with a decision scenario. Although imbedded in a medical context, no medical content knowledge was necessary to complete the assessment tool (Table 2).

Table 2

Instrument Overview

Scale	Maximum Points attainable
Logical reasoning	13
Willingness to learn	60
Epistemological beliefs	
Justification by authority	18
Justification by community	18
Justification by sources	18
Reflexivity of knowledge	18
Personal justification	18
Certainty of knowledge	18
Statistical literacy	30
SRA	
Evidence evaluation (EE)	10
Drawing conclusions (DC)	60

3.2.2.1. Statistical Literacy

For the objective assessment of SL, a broad spectrum from basic to advanced levels was inquired. A total of thirteen items from three different validated tests were used (Appendix A). Duplicate items were excluded and those testing for factual knowledge, so that all three levels described by Watson (1997), including risk literacy, statistical concepts, and the interpretation of statistical data, were covered. All items were weighted for difficulty in terms of the three levels of Watson (1997). Items were adapted and translated into German from the Berlin Numeracy Test (Cokely, 2012, 4 Items; Cronbach's $\alpha = 0.59$), the Obstetrician-Gynecologist Statistical Literacy Questionnaire (Anderson et al., 2014, 3 Items; Cronbach's $\alpha = .53$) and the Medical Data Interpretation Test (Schwartz et al., 2005, 6 Items, Cronbach's $\alpha = .71$). All internal consistency values are presented as reported by the respective authors.

3.2.2.2. SRA

In terms of SRA, this study focused on two out of the eight epistemological activities described in the framework by Fischer et al. (2014): evidence evaluation (EE) and drawing conclusions (DC), which were measured with a decision scenario in medical context (Appendix B).

In the scenario, the participants were asked to put themselves in the role of a General Practitioner (GP) and were asked by a patient for an opinion on additional naturopathic treatment. Initially, the participant made a first decision for or against the treatment on a 6-point Likert scale (1 = rather no to 6 = rather yes). Hereafter, authentic pieces of evidence, presented in short articles on the specific naturopathic treatment, were offered. These evidences were taken from the Apotheken Umschau (Simon, 2016), a pharmaceutical brochure (adapted from: Dr. Wilmar Schwabe GmbH & Co. KG 03/2011, Crataegutt novo 450mg – Erste Wahl bei ersten Beschwerden), the *Ärzteblatt* (Ärzteblatt, 2017), and the *Ärztezeitung* (Meissner, 2017).

Participants indicated no impediment regarding the medical context. As the scenario's focus was to measure SRA, it included little statistical content and stayed on the basic level of SL in the concept of Ben-Zvi and Garfield (2004).

Regarding the evaluation, a denomination of scientific value was assigned to each of these four pieces of evidence (1, 2, 4, and 3 respectively, with 1 representing the lowest value). This rating was independently validated by an expert rating of 71 practicing physicians (Part II). The participants then evaluated each piece of evidence regarding scientific quality, evidence strength, and relevance for the present situation on a 6-point Likert scale based on the QUESTS criteria (Harden, 1999). Based on these Likert-values, the scientific value was calculated and a rating of the participants for each evidence was created. In a final step, it was compared to an expert rating. Absolute differences between the participant rating and the expert rating were cumulated and recoded into a measure of similarity to the expert rating (EE score). A high overall score thus represented a high similarity (maximum 10 score, Cronbach's α .87). Then, the participants were asked to re-evaluate their initial recommendation for or against the treatment (evidence-based decision).

In a last step, the participants rated the persuasiveness of 20 arguments. All 20 arguments were extracted from the presented evidence and assigned with a level of argument strength (1-4, with 1 representing the lowest argument strength). To build an overall score for drawing conclusions, the participants' Likert ratings for persuasiveness of each of the 20 arguments were compared to the expert rating. As with the EE score, the absolute difference between the participant rating

and the expert rating were recoded into a measure of similarity to the expert rating and cumulated for all 20 arguments. A high DC score represents a high similarity to the expert rating (maximum 60, Cronbach's α .74).

3.2.2.3. Control variables

Three scales were chosen as control variables that have been previously described to be intertwined with SL and SRA. The first scale is logical reasoning (Schneewind & Graf, 1998; 13 Items, 1 correct out of 3, Cronbach's α = .79), which correlated well with basic intelligence and was found to be related to SL and numeracy (Lipkus et al., 2001). Furthermore, the achievement motivation could, together with a positive validation of (statistical) skills as described by Griffith et al. (2012), contribute to better scores. Thus, the scale willingness to learn by Schuler and Prochaska (2002) was included (10 Items, 6-point Likert scale, Cronbach's α = .68). Additionally, epistemological beliefs (EB) of the participants have been described in the literature as key elements of scientific skills (Klopp & Stark, 2016, short version, 18 items, 6-point Likert scale, Cronbach's α = .70).

3.2.3. Procedure

The survey was conducted with LamaPoll (https://www.lamapoll.de/, accessed June 16, 2020), a survey tool optimized for mobile applications. After a short introductory text, data use and confidentiality agreement, participants entered a unique personal code, effectively rendering the data anonymous but retrievable. The average test duration was 45 minutes. Participants were advised to refrain from using auxiliaries. Participants were invited via internal university mailing lists, social networks, personal contacts, and existing contact lists of the local testing lab. Participation was voluntary and a small monetary compensation was awarded. The study was approved by the ethics committee of the Medical Faculty of LMU University (approval reference no. 527-16).

3.2.4. Statistical Procedures

The statistical analysis was performed with IBM SPSS 25. Descriptive and frequency data were computed for primary analysis and Cronbach's alpha for reliability of scales. Extensive outlier analyses were conducted and all required prerequisites for statistical analyses were tested. One- and two-factorial ANOVAs were calculated to assess influences of study progress and study domain on the dependent variables SL, EE score, and DC score. Linear regression models were calculated to assess the association of SL and SRA skills under consideration of the control variables. *P* values less than .05 were considered statistically significant.

3.3. Results

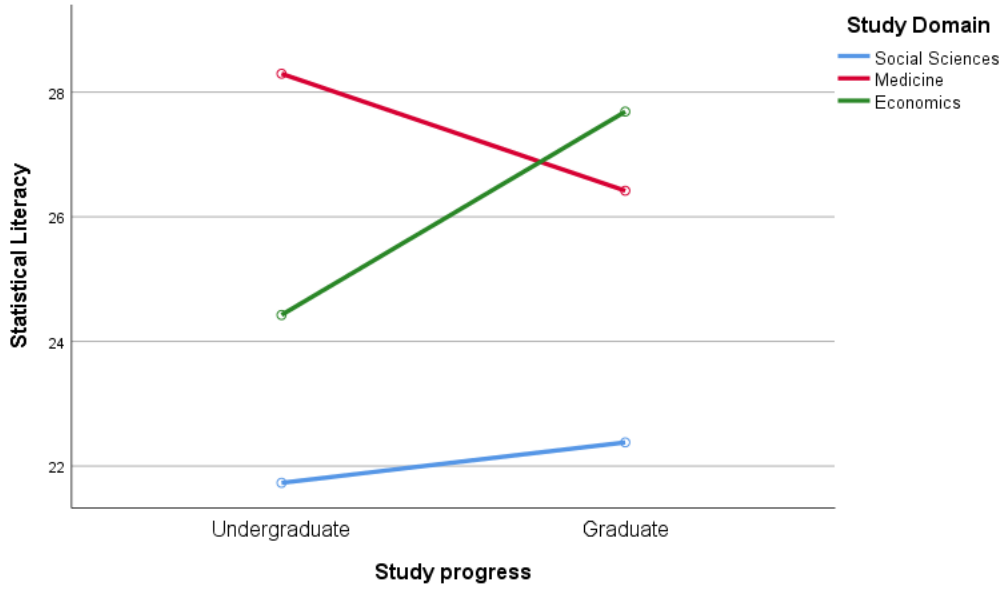
In total, N = 212 complete cases were included (Table 1). For 10 participants, missing values in the EE score were imputed from the average of that specific item. The entire data set was checked for univariate outliers, skewness and kurtosis for all variables was within the ±2 range (Tabachnick & Fidell, 2001). All items of interest were normally distributed and homoscedastic, so that the prerequisites for ANOVA were fulfilled.

3.1. Statistical Literacy

The impact of study domain and study progress on SL (hypotheses 1a-c) were examined with a two-factorial ANOVA. It revealed a significant main effect of study domain: $F(2,211)$ = 14.96, $p < .001$, partial η^2 = .13, with social science students scoring significantly lower than students in economics and medicine ($M_{\text{Social Sciences}}$ = 22.16, SD = 6.45, M_{Medicine} = 27.37, SD = 6.14, and $M_{\text{Economics}}$ = 25.49, SD = 5.57), disproving hypothesis 1a. Pairwise comparisons revealed a significant interaction effect between study domain and study progress for medical and economics students: $F(2,135)$ = 5.53, p = .02, partial η^2 = .04 (Figure 4). Overall, there were differences in SL based on study progress, but none were significant, refuting hypothesis 1b.

Figure 4

SL Score Student Sample

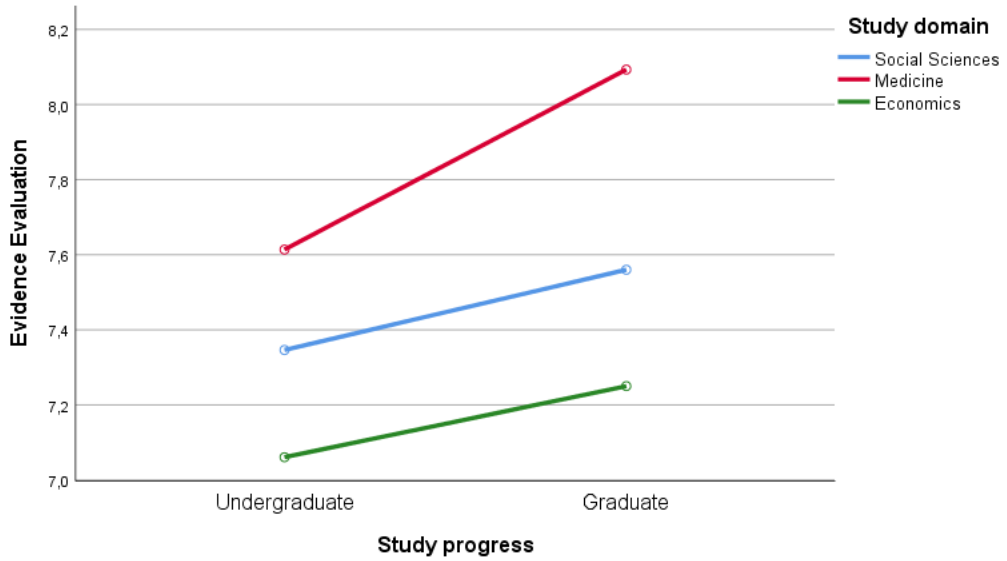


3.2. SRA

When examining the impact of study domain and study progress on SRA skills, EE skills were analyzed first. A two factorial ANOVA revealed no significant effects, $F(5,206) = 1.45$, $p = .21$, disproving hypotheses 2a-b. However, pairwise comparisons revealed a significant difference between medical students (higher score) and economics students ($F(3,135) = 4.74$, $p = .031$, partial $\eta^2 = .35$ (Figure 5).

Figure 5

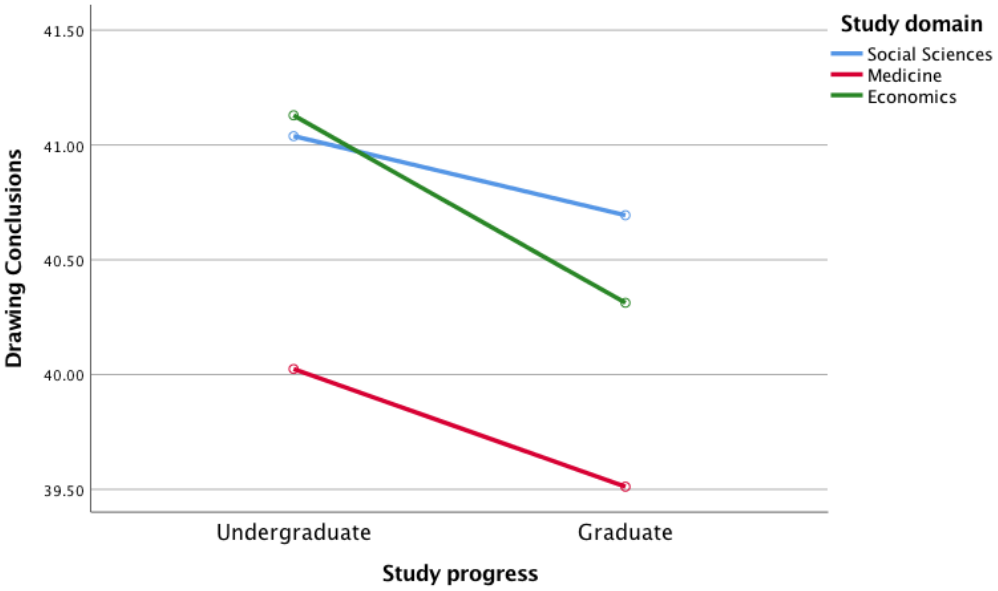
EE Score Student Sample



The participants had to rate 20 arguments extracted from the four pieces of evidence (DC-Score). A two-factorial ANOVA revealed no significant effects: $F(5,206) = 0.887$, $p = .491$ (Figure 6), disproving hypotheses 2a-c.

Figure 6

DC Score Student Sample



EE-Score and DC-Score were not correlated, $r(213) = -.063$. No SRA skill correlated with SL, EE-Score: $r(213) = .072$, $p = .30$; DC-Score: $r(207) = -.072$, $p = .31$. Before and after reading the evidences, participants had to decide for or against the recommendation of the naturopathic treatment. Compared to the first decision, significantly more participants recommended the additional treatment in the second decision: $t(212) = 9.34$, $p < .001$; $M_1 = 3.52$ ($SD_1 = 1.56$); $M_2 = 4.32$ ($SD_2 = 1.26$) and no difference was observed between undergraduate and graduate students in the respective domain (Table 3).

Table 3

Comparison of First and Second Decision

		First decision			Second decision		
		Mean	SD	p-value	Mean	SD	p-value
Medicine	Undergraduate	3.73	1.47	.162	4.50	1.15	.044
	Graduate	3.44	1.62		3.79	1.46	
Social	Undergraduate	3.58	1.33	.031	4.46	.99	.300
Science	Graduate	3.28	1.65		4.20	1.25	
Economics	Undergraduate	3.45	1.70	.100	4.55	1.35	.009
	Graduate	3.81	1.52		4.88	.62	
All		3.52	1.56		4.32	1.26	<.001

3.3. Influencing factors of SL and SRA

Furthermore, it was investigated to what extent EB differs between study domains and study progress. A MANOVA with the six EB as dependent variables showed a significant multivariate main effect for study domain, $F(5,211) = 3.030$, $p < .001$, partial $\eta^2 = .083$ (Pillai's trace = .165). Participants from the three study domains differed significantly in the EB *Authority*, $F(2,206) = 4.770$, $p = .009$, partial $\eta^2 = .044$ ($M_{\text{Social Sciences}} = 8.16$, $SD = 3.10$; $M_{\text{Medicine}} = 9.84$, $SD = 3.11$; $M_{\text{Economics}} = 9.82$, $SD = 2.76$) and the EB *Certainty of Knowledge*, $F(2,206) = 8.553$, $p < .001$, partial $\eta^2 = .077$ ($M_{\text{Social Sciences}} = 9.91$, $SD = 3.30$; $M_{\text{Medicine}} = 11.49$, $SD = 3.61$; $M_{\text{Economics}} = 12.73$, $SD = 3.45$). Duncan post-hoc analyses showed for *Authority* that students from social sciences differed from the other two domains (two subsets) and all three groups differed from each other for *Certainty of Knowledge* (three subsets). Further, there was a significant interaction effect for EB *Justification by Community*, $F(2,206) = 5.597$, $p = .004$, partial $\eta^2 = .052$. Duncan post-hoc analyses revealed that social science and economics students differed from medical students, depending on their study progress (two subsets), $M_{\text{Social Sciences}} = 10.25$, $SD = 3.22$, $M_{\text{Medicine}} = 11.53$, $SD = 2.78$, and $M_{\text{Economics}} = 10.12$, $SD = 3.23$.

Correlations between control variables and SL and SRA are reported in Table 4.

Table 4

Correlation between Control Variables, SL and SRA

	EB 1	EB2	EB3	EB4	EB5	EB6	Willingness to learn	Logical reasoning
EE	-.17*	-.07	.05	-.04	-.17*	.12	.05	.07
DC	-.17*	-.07	-.01	.04	-.01	.232**	.172*	-.06
SL	-.06	.261**	-.03	.20**	.07	.17*	.14*	.52**

Note: EB 1 = personal justification, EB 2 = justification through authority, EB 3 = justification through sources, EB 4 = justification through community, EB 5 = Certainty of knowledge, EB 6 = Reflexivity of knowledge

Furthermore, the role of logical reasoning, EB, and willingness to learn for SL and SRA skills was exploratively analyzed. In linear regression models with stepwise selection, SL was predicted by logical reasoning ($\beta = 1.510 \pm 0.456$, $p < .001$), EB *Authority* ($\beta = 0.259 \pm 0.136$, $p = .021$), and EB *Reflexivity of Knowledge* ($\beta = 0.584 \pm 0.156$, $p = .006$), ($R^2 = .34$, $F(4,208) = 27.01$, $p < .001$). The EE-Score was predicted by the EB *Personal Justification* ($\beta = -0.130 \pm 0.190$, $p = .005$) and EB *Certainty of Knowledge* ($\beta = -0.094 \pm 0.188$, $p = .005$), ($R^2 = .07$, $F(2,210) = 7.32$, $p = .001$). The DC-Score was predicted by EB *Reflexivity of Knowledge* ($\beta = -0.584 \pm 0.234$, $p = .001$), $R^2 = .09$ $F(2,204) = 9.86$, $p = .001$). Willingness to learn and other EB were no significant predictors.

3.4. Summary

The cross-sectional design allowed to compare the performance of undergraduate students with their graduate peers. However, drawing conclusions about development of skills over time is limited. Medical students receive better SL-Scores than social sciences students and are on equal footing with economic students. However, graduate medical students were superseded by their undergraduate peers. Hypotheses 1a-b were thus falsified.

Regarding SRA, students from all three domains are nearly equally skilled in both evidence evaluation and drawing conclusions, also rejecting hypotheses 2a-b. Medical graduate students in particular changed their opinion the least between the first, intuitive decision and the second, evidence-based decision, being rather skeptical of the additional naturopathic treatment compared to the other two study domains. Participants from the three study domains differed significantly in the EB *Authority* and the EB *Certainty of Knowledge*. The evidence provided seem to have thus a different impact on the students from the three domains.

Examining the relationship of SL and SRA, no correlation was observed between SL and SRA scores. SL, EE and DC were predicted by different epistemological beliefs. Logical reasoning was a predictor for SL, but not for SRA.

Taken together, a more advanced study phase or a certain domain alone did not provide a sufficient indicator for advanced SL and SRA skills, suggesting that these skills are not automatically developed over the course of university studies. Rather, they could be considered as dependent on domain-specific characteristics, like curricula, appreciation of these skills and epistemological beliefs.

4. SL & SRA in Physicians – Part II

4.1. Introduction

Statistical literacy and SRA seem to be better developed in physicians in comparison to medical students (see theoretical background). Epstein et al. (2018) suggested that this improvement does not necessarily happen within formal medical education. However, the question remains how, where, and when this improvement arises. The second part of this dissertation aims at answering research questions 2 and 3.

4.2. Methods

4.2.1. Design and Sample

This experiment followed a quasi-experimental, causal-comparative design with the two dependent variables SL and SRA. In total, $N = 71$ German-speaking physicians (31 females, 34 males) were included (Table 5).

Table 5

Description of Physician Sample

Variable	Options	Frequency	Percentage
Gender	Male	34	47.9%
	female	31	43.7%
	No answer	6	8.4%
German mother tongue	Yes	68	95.8%
	No	3	4.2%
	NA	5	7.0%
MD in Germany	Yes	68	95.8%
	No	3	4.2%
MD-Thesis	Yes	58	81.7%
	Currently working on	9	12.7%
	No thesis or skipped	4	5.6%
Academic qualification	Habilitation	12	16.9%
	Professorship	7	9.9
Ever worked as researcher	Yes	36	50.7%
	No	35	49.3%
Working Environment	Hospital	2	2.8%
	Out-patient care	43	60.6%
	Research	16	22.5%
	NA	10	14.1%

They came from different work settings and locations, namely hospital (*N* = 43), outpatient sector (*N* = 2) and research (*N* = 16). A MD-thesis, a scientific work as optional part of medical studies, was completed by 58 participants and 9 were currently working on it. This sample can be considered representative regarding scientific experience. The mean age of participants was 40 years (*SD* = 9.59, *range* = 26 - 65)

4.2.2. Test Instrument

The same measurement tool was used for the assessment of SL and SRA skills as applied to the student sample in Part I. Reliability of the SL scale was .82 (Cronbach's α) in this sample with a maximum score of 30 points. A section for the assessment of subjective measures was added. The subjective numeracy test (Fagerlin et al., 2007) was applied to evaluate numeracy (8 Items, Cronbach's α = .84,). For the assessment of subjective SL, six items were newly developed (Likert scale 1-6; Cronbach's α = .90, maximum score 36/36). Scores were calculated for subjective numeracy and subjective SL by adding up the Likert values respectively.

For this section, an extensive ensemble of demographic items was added focusing on the work history and environment (hospital, out-patient care, research) of the participating physicians. Questions were adapted from a study by Epstein et al. (2018) and comprised multiple choice items, some with the opportunity to fill in additional free text; Five items on the MD-thesis (qualifications fostered meanwhile), three items on the professional career, two items on the publication record (type of authorship, number of publications), and three items on the current job description (Epstein et al., 2018). These items on relevant demographic factors were piloted with ten medical students from the LMU Munich.

4.2.3. Procedure and Analyses

Analogously to the student study sample, the participants could complete the survey online with LamaPoll. Additionally, participants could complete the survey with pen and paper (return rates: 16.5% online and 66.7% pen and paper). Average duration was 45 minutes, comparable to the time required by the students in Part I. Participants were invited via mailing lists and personal contacts.

Statistical analysis was also performed with IBM SPSS 25, using the same analytical tools and operations as in Part I. Additionally, data in natural verbal language, e.g. free text in demography section, underwent thematic analysis to extract common themes.

4.3. Results

There were *N* = 71 completed questionnaires included in the analysis (Table 5). For 13 participants, missing values in the evaluation of the 20 arguments from the average of the respective item were imputed. Analogously to Part I, the entire data set was checked for univariate outliers. The prerequisites for ANOVA were fulfilled, all variables were normally distributed and homoscedastic.

4.3.1. SL and SRA

The 71 physicians' average score in SL was *M* = 17.58, *SD* = 6.92, with a range of 5 to 30 out of 30 attainable points. These scores correlated significantly with their subjective SL, $r(71) = .34$, $p = .004$) and also their subjective numeracy, $r(71) = .33$, $p = .004$).

On average, physicians evaluated the evidences concordantly with the preassigned evaluation, EE-Score *M* = 7.75, *SD* = 1.85. The ratings for argument quality were also in line with the preassigned rating, DC-Score *M* = 37.20, *SD* = 5.35. Statistical literacy and DC were significantly inversely correlated, DC-Score $r(71) = -.272$, $p = .022$. However, no correlation was found between SL and EE, $r(71) = .198$, $p = .098$, nor EE and DC, $r(71) = .138$, $p = .256$.

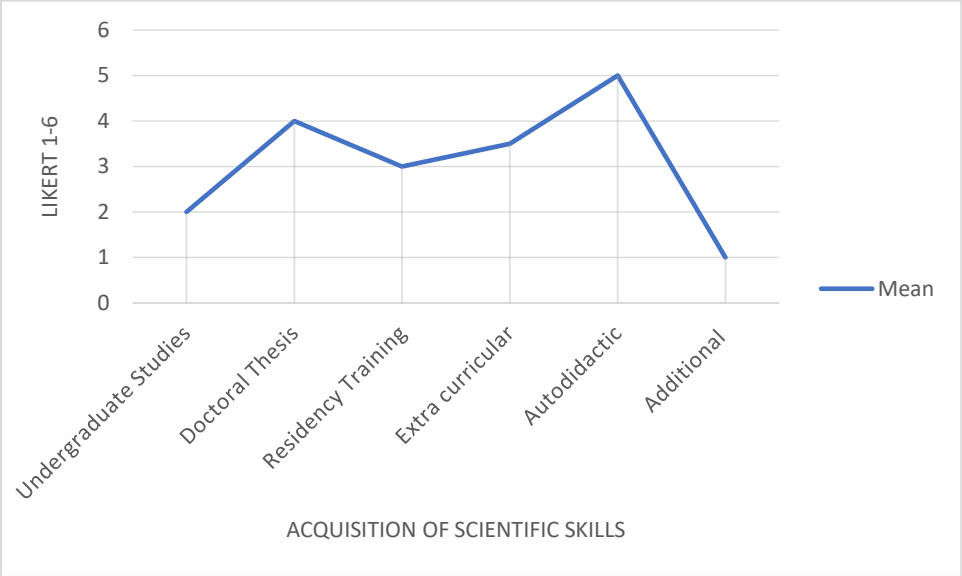
4.3.2. Education & Development in Physicians

Furthermore, it was explored how, where, and when physicians developed SL and SRA skills. Significantly more physicians specified that their skills had been developed autodidactically, *M* = 4.78, *SD* = 1.13, Likert 1-6 scale, rather than to have happened during their studies, *M* = 2.31, *SD* = 1.46, Likert 1-6 scale, $t(71) = -9.915$, $p < .001$, or in extracurricular activities, *M* = 3.34, *SD* = 1.87, Likert 1-6 scale, $t(71) = 4.673$, $p < .001$, (Figure 7). A free-text box offered the opportunity to indicate other learning occasions. Physicians added, *inter alia*, massive open online courses, postgraduate studies (such as master's degrees), workshops, and learning through peer reviews and feedback (Figure 7).

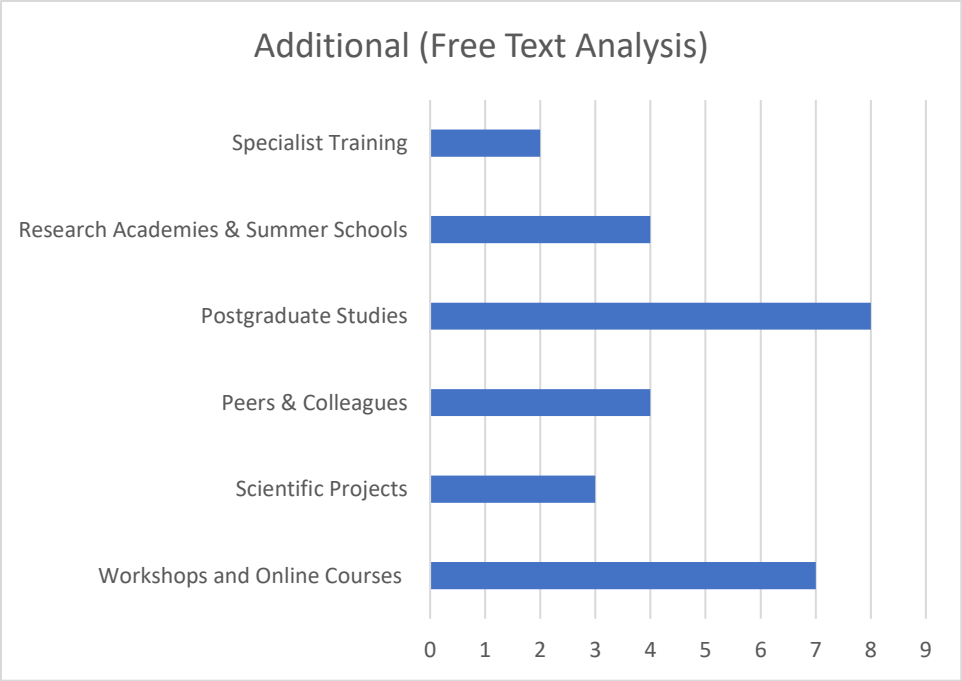
Figure 7

Acquisition of Scientific Skills

A



B



Note: B constitutes the free text analysis of the option *additional* in A

Regarding the MD-thesis, a univariate ANOVA showed that having completed or working on a MD-thesis had an effect on the scoring in EE, $F(3,71) = 10.494$, $p < .001$, partial $\eta^2 = .320$, but not on DC, $F(3,71) = 1.133$, $p = .342$, nor SL, $F(3,71) = 1.812$, $p = .153$, partial $\eta^2 = .075$. The fostering of critical evaluation of study results presented by other researchers during the

preparation of the MD-thesis positively correlated with the scoring in SL, $r(71) = .271$, $p = .033$.

Regarding the postgraduate phase, a one-factorial ANOVA showed a significant main effect of having worked in research on SL, $F(1,70) = 12.737$, $p = .001$, partial $\eta^2 = .156$. The type of authorship, $F(5,71) = 3.886$, $p = .004$, partial $\eta^2 = .230$, and the time spent in research, $F(23,71) = 2.262$, $p = .009$, partial $\eta^2 = .525$, were significantly associated with better SL, as well as the number of publications, $r(71) = .36$, $p = .002$.

Regarding SRA, linear regression models revealed that the corresponding score in EE increased by $\beta = .314 \pm .150$, $p = .041$, when the Likert value of the MD-thesis supervisor's content-related support was increased by one point. Additionally, the form of the MD-thesis (experimental, clinical, empirical, statistical, or literature review) was associated with EE, with experimental and clinical design being positively related to EE skills, $\beta = -.380 \pm .154$, $p = .016$, $R^2 = .187$, $F(1,59) = 4.353$, $p = .041$. DC was higher when participants indicated to have already worked in research, $\beta = 3.355 \pm 1.229$, $p = .008$, $R^2 = .314$ $F(1,68) = 7.448$, $p = .008$.

4.4. Summary

Statistical literacy of physicians was at a medium level, correlated well with subjective measures and was thus comparable to other higher educated study groups. Regarding SRA skills, the picture was not consistent. While evidence evaluation was found to be at a medium-advanced level, DC score was lower and thus less in accordance with the preassigned ratings.

Most participants indicated to have acquired scientific skills outside of formal medical education in an autodidactic manner, in postgraduate studies or extracurricular activities. For enhanced SL, having worked in research (the longer the better) and published papers (preferably as first author, and the more the better) were key factors. A higher EE score was yielded when having completed a research project with, favorably, experimental or clinical design and content-related support was provided by the supervisor.

Therefore, it seems important for the development of SL and SRA that medical students are involved in the process of research and writing papers already during their doctoral thesis. Furthermore, formal training of research and statistical skills should be continued during residency.

5. SL & SRA in the medical domain – Part III

5.1. Introduction

As the same measurement tool was applied in Part I and Part II, this part proceeds with a direct comparison of the two study groups. A medical sample was created, combining the assessment of medical students (undergraduate and graduate) and physicians. This could offer additional insights into SL and SRA in the medical domain and domain-specific epistemological beliefs, willingness to learn and logical reasoning skills. Thus, it will shed light on the relationship of SL and SRA and provide additional information on how to better foster these skills. This part hence addresses research question 2.2., whether SL and SRA skills of physicians are superior to those of medical students and contributes to the investigation of the relationship of SL and SRA, tackling the fourth research question.

5.2. Methods

For the comparison of medical students and physicians, a new dataset was created from the samples I and II, described in Part I and II. It comprises a total of $N = 157$ participants, $N = 43$ undergraduate medical students, $N = 44$ graduate medical students and $N = 70$ physicians (medical sample). The statistical analysis was conducted analogously to Part I and II with SPSS 25 with the same analysis techniques.

5.3. Results

The entire data set was again checked for univariate outliers, skewness and kurtosis (for all variables within the ±2 range). The prerequisites for ANOVA were fulfilled for all relevant items. Demographics of the study sample have already been described in Part I and II, respectively.

5.3.1. Statistical Literacy

Statistical literacy did not differ significantly across the different stages of medical education, $F(2, 156) = 1.567$, $p = .212$, $M_{undergraduate} = 15.23 \pm 4.26$, $M_{graduate} = 13.43 \pm 4.84$, $M_{physicians} = 15.03 \pm 6.20$). However, by trend, medical students in their second phase attained a lower score in SL compared to students in their early years of studying as well as physicians.

While the participants did not differ in their self-rated numeracy, $F(2, 156) = 1.742$, $p = .179$, subjective SL varied significantly, $F(2, 156) = 4.503$, $p = .013$, partial $\eta^2 = .055$, $M_{undergraduate} = 19.33 \pm 6.90$, $M_{graduate} = 17.20 \pm 6.55$, $M_{physicians} = 21.00 \pm 6.41$.

5.3.2. Scientific Reasoning and Argumentation

In terms of SRA, there was no significant difference observed between undergraduate, graduate medical students and physicians regarding evidence evaluation, $F(2, 154) = .791$, $p = .455$. All means are depicted in Table 6. However, they differed significantly regarding drawing conclusions, $F(2, 151) = 6.638$, $p = .002$. The Waller-Duncan Test for homogenous subgroups revealed that physicians differ significantly from the student sample (two subsets).

Table 6

SRA Comparison

		Evidence Evaluation		Drawing conclusions	
		Mean	SD	Mean	SD
Students	Undergraduate	2.37	1.20	40.02	3.97
	Graduate	1.93	1.208	39.51	3.48
Physicians		2.27	1.86	37.11	5.34
All		2.20	1.74	38.56	4.70

Before and after reading the evidences, participants had to decide for or against the recommendation of the naturopathic treatment. While medical students differed significantly in their recommendations, physicians did not change their opinion after having read the evidences, physicians: t (71) = -.922, $p = .359$; $M_1 = 2.75$ ($SD_1 = 1.57$); $M_2 = 2.87$ ($SD_2 = 1.50$); students: t (87) = 3.50, $p = .001$; $M_1 = 3.59$ ($SD_1 = 1.54$); $M_2 = 4.15$ ($SD_2 = 1.46$), Table 7.

Table 7

Comparison of First and Second Decision

		First Decision		Second Decision		
		Mean	SD	Mean	SD	p – value*
Students	Undergraduate	3.73	1.47	4.50	1.15	
	Graduate	3.44	1.62	3.79	1.46	
	All	3.59	1.54	4.15	1.35	.001
Physicians		2.73	1.58	2.86	1.50	.359

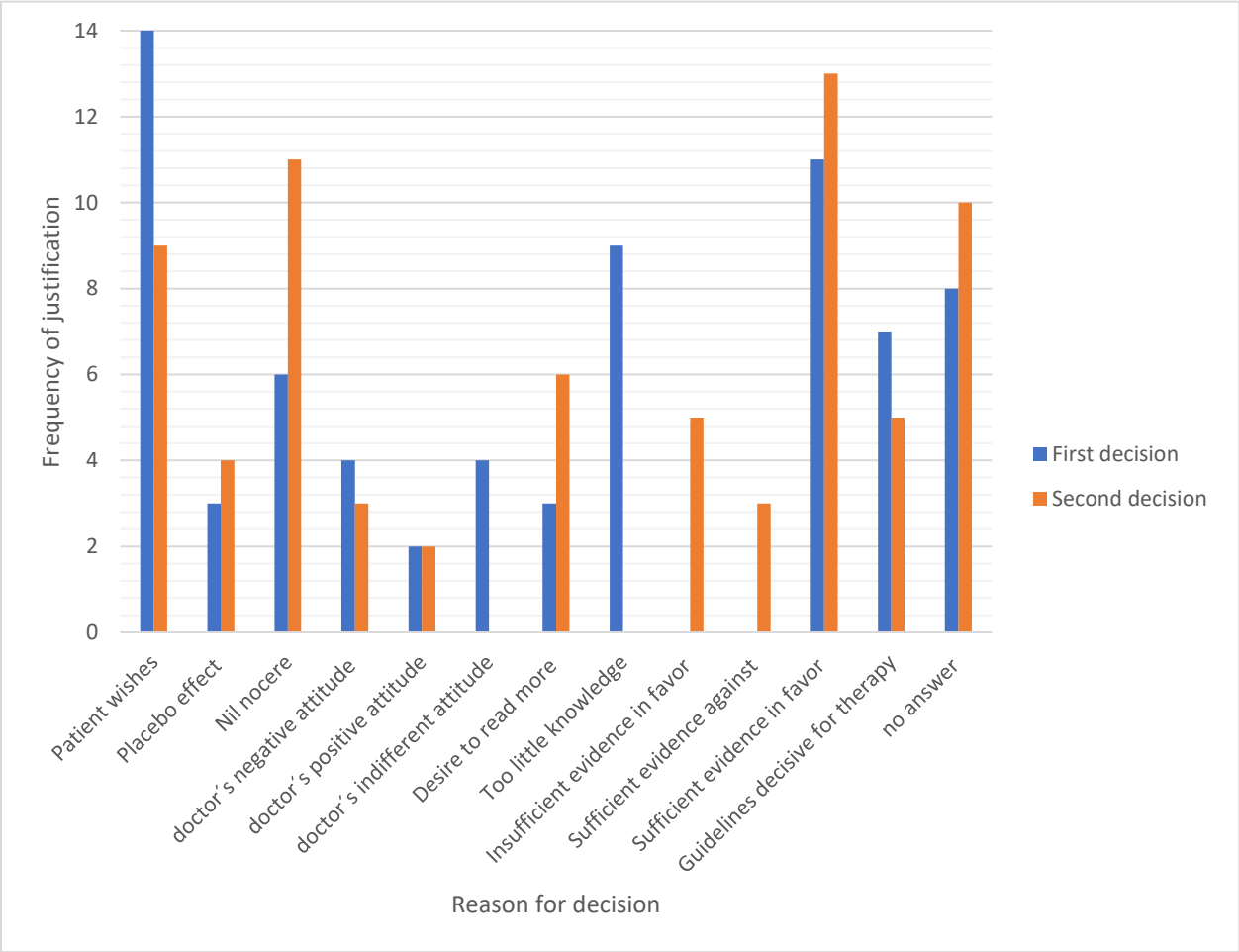
Note: *compares medical students´ first and second decision, and physicians´ first and second decision

A one factorial ANOVA also showed significant differences between students and physicians in their first and second decision, $F_{first}(2,155) = 8,289$, $p < .001$, $F_{second}(2,155) = 17.785$, $p < .001$. In the physician sample, an extensive analysis of the free-text justification was conducted revealing the following principal reasons for their recommendation: patient wishes and the

principal of doing no harm (*nil nocere*) among sufficient evidence provided in favor of naturopathic treatment for the second decision (Figure 8).

Figure 8

First and Second Decision



Note: Free Text Analysis

5.3.3. Influencing Factors

A small set of covariables and demographics was collected in both study samples. These were for example achievement motivation, logical reasoning, epistemological beliefs and factors about the MD-thesis.

In terms of achievement motivation and logical reasoning, physicians and medical students were found to be very similar. Achievement motivation: $F(2, 156) = 2.606$, $p = .077$, $M_{undergraduate} = 40.42 \pm 6.437$, $M_{graduate} = 41.39 \pm 6.322$, $M_{physicians} = 43.11 \pm 6.280$. Logical reasoning: $F(2, 156) = 1.423$, $p = .244$, $M_{undergraduate} = 11.74 \pm 1.217$, $M_{graduate} = 11.27 \pm 1.590$, $M_{physicians} = 11.29 \pm 1.669$.

In terms of epistemological beliefs, physicians and medical students did not differ in EB *Personal Justification*, $F(2, 156) = .551$, $p = .577$, in EB *Justification by Sources*, $F(2, 156) = .612$, $p = .543$, in EB *Certainty of knowledge*, $F(2, 156) = .480$, $p = .620$, and EB *Reflexivity of knowledge*, $F(2, 156) = .151$, $p = .860$. Physicians and medical students had diverging epistemological beliefs regarding EB *Authority*, $F(2, 156) = 11.343$, $p < .001$, partial $\eta^2 = .128$, $M_{undergraduate} = 10.00 \pm 3.200$, $M_{graduate} = 9.68 \pm 3.041$, $M_{physicians} = 7.47 \pm 3.101$ and EB *Justification by Community*, $F(2, 156) = 3.079$, $p = .049$, partial $\eta^2 = .038$, $M_{undergraduate} = 11.84 \pm 2.768$, $M_{graduate} = 11.23 \pm 2.794$, $M_{physicians} = 10.50 \pm 2.878$.

In this sample, there were $N = 60$ participants with a completed MD-thesis, $N = 93$ were working on it, $N = 3$ were neither working on it nor had completed a MD-thesis. A single participated cancelled their project. Due to this very homogenous picture, there was no significant difference observed in SL with regard to the MD-thesis, $F(3, 156) = 2.510$, $p = .061$, $M_{thesis} = 24.97 \pm 6.454$, $M_{inprogress} = 23.78 \pm 5.084$, $M_{nothesis} = 19.33 \pm 3.055$, $M_{cancelled} = 35.00$.

5.3.4. Model Building

A significant correlation of DC and EE was found in this dataset, $r(154) = -.160$, $p = .046$. Statistical literacy score, however, was associated with neither EE, $r(157) = .064$, $p = .429$ nor with DC, $r(154) = -.124$, $p = .125$.

In linear regression models with stepwise selection, SL was predicted by logical reasoning, $\beta = 1.138 \pm 0.263$, $p < .001$, and EB *Reflexivity of Knowledge*, $\beta = 0.642 \pm 0.242$, $p = .009$, $R^2 = .14$, $F(1,151) = 7.09$, $p = .009$.

The EB *Certainty of knowledge* predicted the EE-Score, $\beta = 0.097 \pm 0.038$, $p = .013$, $R^2 = .04$, $F(1,152) = 6.39$, $p = .013$) and the DC-Score was modeled by EB *Authority*, $\beta = 0.303 \pm 0.110$, $p = .007$, age, $\beta = -.093 \pm 0.033$, $p = .006$, evidence evaluation skills, $\beta = -.509 \pm 0.199$, $p = .011$, and the EB *Reflexivity of Knowledge*, $\beta = -0.517 \pm 0.209$, $p = .014$), $R^2 = .193$, $F(1,149) = 6.146$, $p = .014$.

Willingness to learn and other EB were no significant predictors.

5.4. Summary

In the third part of this dissertation, the two study samples in the medical domain were directly compared. The application of the same measurement tool in physicians and medical students allowed to gain insights in these skills in the realm of a German university environment.

Statistical literacy as well as evidence evaluation scores are very similar in physicians and medical students of all study phases. Drawing conclusions in physicians is not superior to that of students leading to a rejection of the hypotheses 3a-b. Regarding the treatment recommendations, medical graduate students constituted the most skeptical of all students, however, physicians did not change their recommendation at all. Decision making in physicians might thus be subjected to other factors, being discussed in section 6.5. Medical students and physicians do not differ in their achievement motivation or their logical reasoning skills but display differences in the EB *Justification by Community* and A*uthority*.

6. Discussion

6.1. Introduction

The aim of this dissertation was to gain insights in SL and its relevance for and relationship with SRA in medical students and physicians. It should contribute to the discussion how to foster these skills in lifelong learning of physicians. Therefore, a unique measurement tool was created to assess both, SL in the comprehensive definition of Watson (1997) and SRA, focusing on the two epistemological activities EE and DC (Fischer et al., 2014). First, SL and SRA were examined in students. This cross-domain measurement compared medical students to that of economics and social sciences across two study phases (research question (RQ1), Part I). Second, these skills were analyzed in physicians (RQ2, Part II). In the same sample, additional demographic items provided insights in how, when and where these skills were acquired (RQ3). Finally, influencing factors, such as logical reasoning, epistemological beliefs, and willingness to learn were exploratively assessed in order to contribute to the goal of modelling the relationship of SL and SRA with regard to scientific reasoning (RQ4).

6.2. SL and SRA in Students (RQ 1)

The first research question examined the impact of study domain and study progress on SL (RQ 1.1.). It was hypothesized that students in economics and social sciences attain higher scores in comparison to medical students due to their extensive methodological curriculum. This could not be confirmed. Students of medicine and economics superseded social sciences students in both study phases. Regarding the comparison of first and second study phase, i.e. undergraduate and graduate, medical students were found to receive lower test results in their graduate phase. Contrastingly, economics students were observed to have better SL skills in their graduate phase.

These findings are in line with two threads of research. First, in the early years of university studies, knowledge from school might be more present than in later phases. Since a lot of effort has been done to enhance the learning of statistics in school (Budgett & Rose, 2017) and it might not be adequately fostered during university studies, SL could thus decline over time. Second, attitudes towards statistics and its necessity in professional life were considered to have a greater impact on SL than actual declarative knowledge about statistics (Griffith et al., 2012; Williams, Payne, Hodgkinson, & Poade, 2008). Studies suggested that medical students consider statistics less relevant for their later practice (Altman & Bland, 1991; Gigerenzer, 2002), whereas graduate economic students were found to have a positive attitude towards

statistics (Griffith et al., 2012). Gaissmaier and Gigerenzer (2008) provided evidence that the mere concentration on teaching of statistical techniques does not imply the improvement of SL. Hence, it seems important to not only foster SL adequately during university studies but also to increase the students' awareness of significance SL has in their respective domain and in everyday life.

The second part of the first research question examined the impact of study domain (medicine, social sciences, economics) and study progress (undergraduate, graduate) on SRA skills of students (RQ 1.2.). Economics and social sciences students were hypothesized to attain higher similarity scores in SRA than medical students due to their more extensive research curriculum. This could not be confirmed. Instead, scores in both evidence evaluation and drawing conclusions did not differ across students from all three domains. Furthermore, SRA skills did not change over time. This is contrary to White et al. (2011) who assessed critical thinking skills with the ACTA in biology and chemistry students. They found that these capabilities improved over the course of studies.

Nonetheless, the evidences provided in this study seem to have a different impact on the students from the three domains. This was reflected in the significant difference between their first, intuitive decision and second, evidence-based decision about the naturopathic treatment. Before the evidence was presented, students from all study domains and all study phases were similarly undecided (Table 3). Conversely, the majority of the students changed the decision in favor of the additional neuropathic treatment for their second decision, constituting an important impact of the evidences on the decisions of the participants. Interestingly, on average, medical graduate students changed their opinion the least between the two decisions, remaining more skeptical of the additional naturopathic treatment, compared to the other two study domains. There was a significant difference between undergraduate and graduate medical students regarding their first and second decision. Graduate medical students were the most critical group regarding the additional naturopathic treatment and significantly differed from the group of graduate economics students who formed the most positive group on the treatment. Clark and Forster (2017) argue that “the entry requirements, the expectations of students, and the epistemological frameworks that shape the social sciences, are not necessarily the same as they are in [other] disciplines” (p. 260).

Taken together, the characteristics varying across the three domains might be a reason for a different processing of the presented evidences resulting in different recommendations.

6.3. SL and SRA in Physicians (RQ2)

Part II was designed to answer research questions 2 and 3, focusing on SL and SRA in physicians. Additionally, it should shed light on the questions when and how these skills were developed. Statistical literacy of physicians was found to be mediocre (59%) and well correlated with subjective measures, EE score (77%) at a rather high level and DC score (62%) on a medium level.

Since basic numeracy was not regarded in the scale of SL (Cokely et al., 2012), the test discriminated well, and ceiling effects as observed in other educated samples (Lipkus et al., 2001; Hanoch et al., 2010) were not present. Further studies assessing SL of physicians were designed upon varying theoretical bases. A study focused on knowledge of 18 different statistical tests among pathologists and observed a rather low level of SL (Schmidt et al., 2017). Anderson et al. (2014) did not create an overall score but distinguished between fact, concept, and relation questions and found altering levels of SL. A study with Greek residents also concentrated on knowledge questions and reported a rather low SL (Msaouel et al., 2014). Riegelman and Hoveland (2012) found that residents struggled when critical reflection upon research was required. The results of this dissertation are thus comparable to the literature. Future research should aim at a continuous assessment of SL in physicians with similar instruments on the basis of a clear and continually used theoretical foundation.

6.4. Education and Development (RQ3)

Furthermore, with this sample of German-speaking physicians, the questions how, where, and when physicians developed SL and SRA skills were examined. Participants indicated to have acquired scientific skills mostly in an autodidactic manner, in higher education outside of medical studies (such as master's degrees) or in extracurricular activities.

Statistical literacy was enhanced when critical thinking was fostered within the phase of working on the MD-thesis. It was improved when physicians had been or were currently working in research. Moreover, the number of publications and the type of authorship was positively associated with greater SL scores. These findings are in line with Schmidt et al. (2017), stating that an advanced degree other than MD or statistic courses as positively associated with SL. A study with physicians, residents, and final year medical students in Thailand showed that statistical workshops completed only recently do indeed lead to higher SL scores (Laopaiboon, Lumbiganon, & Walter, 1997). Similar results were presented by White et al. (2011). Veilleux et al. (2017) found that individuals who had college- but not high

school-level coursework in statistics research methods, and psychology attained higher scores than people without them. Thus, it can be established that teaching statistics provides measurable benefits/has considerable benefits. However, White et al. (2011) found that "students' ability to think critically improved over the course of their university studies, which could mean that students with advanced scientific abilities self-select for progressed science training." (p.105).
Furthermore, additional courses are often hard to integrate in medical training, especially postgraduate. More than a third of American Ob-Gyn residents do not receive formal training (Anderson et al., 2013). Additionally, acceptance for interventions overall is limited among neurology residents. These results can most likely be found among residents of all disciplines (Leira, Granner, Torner, Callison, & Adams, 2008).

Regarding SRA skills, advanced evidence evaluation skills were associated with having completed a research project (e.g. MD-thesis). The design of the project should favorably be experimental or clinical. Having content-related support by the supervisor was also positively associated with better EE scores. These findings are in line with the subjective impression of German medical graduates with a MD-thesis ranking their scientific skills higher compared to those working on it (Epstein et al., 2018).

Epstein et al. (2018) found that medical graduates evaluate their own scientific skills after medical school as rather low. In Germany, most of medical graduates regard their research skills as too poorly to conduct research on their own. This is particularly interesting as having already worked in research was associated in this study with enhanced SL and DC scores. The same results were found by Schmidt et al. (2017). In the United States, only 68.1% of medical students in their final year participated in research during medical school and only 42% had (co-)authored a paper submitted for publication. It should be argued that – in order to promote scientific skills in physicians – medical students should be involved in the publishing process and learn how to write papers. There is strong evidence that this might enhance their scientific skills (SL and SRA) in the long run.

6.5. Model Building (RQ4)

This dissertation offered a unique opportunity to examine the relationship between SL and SRA in the medical domain (Figure 3). With the same measurement tool being applied to two different samples, namely medical students and physicians, this setup allowed to model SL and SRA. Guiding research questions for this purpose were (RQ 4.1.): In what manner are statistical

literacy and scientific reasoning and argumentation skills intertwined? (RQ 4.2.): What is the role of other influencing factors such as logical reasoning, EB, and willingness to learn for SL and SRA skills?

6.5.1. Framework

On a very abstract level, two models of thinking have been established: the Structure-of-Intellect Model by Guilford (1967) and the theoretical model established in social sciences (Dietrich et al., 2015). They describe how the *input*, i.e. the exposure in the respective domain, is processed individually (*operation*) in order to produce a certain, measurable *output*. The process, however, is shaped through personal characteristics, such as openness, motivation, confidence, endurance, and intelligence. The input, i.e. knowledge as taught in the respective domains, is thus processed according to the epistemological beliefs and has an impact on the output, such as the measurable skills SL and SRA.

6.5.2. Relationship of SL and SRA

First, it needs to be clarified whether SL and SRA might be seen as two separate concepts or *outputs*, in the sense of Dietrich et al. (2015).

In section 2.3., SL and SRA were depicted as two different yet integral components of scientific reasoning. The results from the student sample (Part I) seemed to underline this assumption. It showed that the scores in statistical literacy, evidence evaluation and drawing conclusions were not correlated.

In the physician sample (Part II), DC and SL were found to be inversely correlated and in the medical sample (Part III), comprising medical students of all phases and physicians, SL and SRA skills were not correlated, supporting the results above. In this sample, a significant correlation of DC and EE was found.

These findings are in line with Opitz et al. (2017) considering SRA not as one single activity but rather as a set of different coordinated skills. Also, in other contextual frameworks, SL has been regarded as a prerequisite for SRA (Watson & Callingham, 2003). In a Dutch community-based study, more numerate participants showed enhanced performance in SRA due to increased evaluation of pros and cons in decision-making and evaluation of judgments (Ghazal, Cokely, & Garcia-Retamero, 2014).

As evidence has not been predominantly presented in numerical or statistical terms, the missing link of SRA and SL was expected. Drawing conclusions displayed an antithetical relationship with SL in the physician sample and with EE in the medical sample, indicating that DC in physicians needs to be further investigated.

6.5.3. Influencing Factors

In a next step, it was examined how the operations in the sense of Dietrich et al. (2015) were influenced. Factors such as logical reasoning, epistemological beliefs (EB), and willingness to learn were considered in this analysis. Regression models found that SL was predicted by logical reasoning, EB *Reflexivity of Knowledge* and in the student sample as well by EB *Authority*. While the EE-Score was predicted by the EB *Personal Justification* and EB *Certainty of Knowledge* in the student sample, the EB *Personal Justification* was no longer a predictor in the medical sample. Regarding the DC-Score, the predictors clearly varied between the medical and the student sample. In the student sample, the DC-Score was only associated with the EB *Reflexivity of Knowledge*, whereas in the medical sample it was modeled by EB *Authority*, EE-Score, age and the EB *Reflexivity of Knowledge* indicating differences in the way conclusions are drawn in the medical field, especially in physicians.

Logical reasoning was strongly associated with SL and is found to be highly correlated with common intelligence tests. This is in line with existing studies that observed positive correlations (Cokely et al., 2012; Martin, Hughes, & Fugelsang, 2017).
Physicians and medical students display very similar achievement motivation and logical reasoning scores, indicating that these scales are based on specific traits such as intelligence, ambition and grid (Lipkus et al., 2001; Schneewind & Graf, 1998; Schuler & Fintrup, 2002; Schuler & Prochaska, 2003). It could be hypothesized that they belong to a set of characteristics that are not subjected to major changes over time and that people with the same reasoning skills and achievement motivation self-select in medicine.

6.5.4. Epistemological Beliefs

In general, epistemological beliefs are embedded in culture, achievement motivation (Urhahne, 2006), and self-regulated learning (Schommer-Aikins, 2004). They are known to shape school achievement and choice of college majors (Trautwein & Ludtke, 2007) and have been observed to differ across domains. A study analyzing textbooks from various study courses found that scientific and mathematical disciplines tend to present facts without references, while psychological textbooks more often refer to sources (Smyth, 2001). This could potentially shape epistemological beliefs.

In this dissertation, the EB *Authority* was assessed by three items that essentially measure how much confidence and credibility is granted to scientists and facts derived from scientific research. The adaption of knowledge presented by authorities (like scientists, literature,

politicians or experts) in an unreflected and unfiltered manner was described as naïve (Schommer, 1990). A person with an advanced epistemological conviction would rather critically agree with experts. A lower score in the scale EB *Authority* applied in this dissertation indicates a more critical stance towards authorities. Social sciences students differed significantly from medical or economics students with showing a more critical stance towards authorities (Klopp & Stark, 2016). Naturally, with gaining more experience and advancing in their careers, medical students and physicians gradually take on a more analytical approach towards authorities, resulting in diverging EB *Authority*.

The EB *Certainty of Knowledge* indicates a stance towards scientific knowledge as being less (more) certain and less (more) of a *final truth* depending on a higher (lower) score. In the present analysis, EB *Certainty of Knowledge* was the most important predictor of evidence evaluation skills. This EB could thus shape the ability to integrate contradicting evidences into a meaningful conclusion (White et al., 2011). Students from all three domains differed in the EB *Certainty of Knowledge*, with social sciences students showing the lowest score, followed by medical students and economics students with the highest scores. There was no difference between medical students and physicians observed.

In the medical sample, EE-Score was no longer predicted by the combination of EB *Personal Justification* and EB *Certainty of Knowledge*, but by EB *Certainty of knowledge* alone. A lower score in the EB *Personal Justification* indicates a stance towards scientific insights being less driven by personal views of the researcher. Interestingly, the difference observed between the student and the medical sample regarding this association could indicate that the EB *Personal Justifications* are being sidelined by the development of other EB, critical thinking and EE skills. There were no differences regarding the EB *Personal Justification* between the students form different domains, nor between medical students and physicians.

The EB *Reflexivity of Knowledge* was a significant predictor of SL, as well as DC, constituting the only two EB being associated with SL and SRA in both samples. A lower score in the EB *Reflexivity of Knowledge* indicates an attitude regarding knowledge as being more static and unchanging in nature instead of being constantly evolving due to new findings. No differences across domain nor between medical students and physicians were found.

The EB *Justification by Community* stands for a stance towards the need for justification of scientific insights by the scientific community, scientific discussion, and publication. Social sciences and economics students display differences in their EB *Justification by Community* dependent on their study progress. Economic students score lower in the undergraduate phase

than in the graduate phase indicating an increased need for justification of scientific results by the scientific community, while social sciences students score lower in their graduate phase than in their undergraduate phase. Medical students remain approximately the same in both study phases but with an elevated score compared to the other two domains. In physicians, however, this EB was found to be lower. This is in line with Dietrich et al. (2015) who described this decrease over the progression of the career alongside the gradual increase in experience. Although the EB *Justification by Community* was no significant predictor for neither SL nor SRA in this analysis, it demonstrates well that EB do not need to be static, rather they are subject to change throughout university studies and in during medical practice. This was also observed by Diamond and Stylianides (2017) arguing that due to the significant differences in personal epistemologies among students and experts in statistics, a domain-specificity of personal epistemologies is needed.

6.5.5. Drawing Conclusions in Physicians

Recapitulating from Part III, DC-Score was predicted by the EB *Authority*, EE-Score, age and the EB *Reflexivity of Knowledge*, while in the student sample, it was only associated with the EB *Reflexivity of Knowledge*. Furthermore, in the physician sample, DC was characterized by an inverse relationship with SL. It seems that drawing conclusion in physicians differs somehow from that of not only students in general, but also from medical students.

Results from Part II indicated that physicians have the most dissimilar rating in comparison to the expert rating (i.e. lowest DC score). A possible explanation could be the design of the measurement tool. It was intended to compare the participants' rating to an expert rating, i.e. the author of this dissertation and her study group. Their preassigned scientific values were compared to that of the $N = 71$ physicians and found to be very similar. When then applied to the student sample, this measurement worked very well. However, when physicians are the primary target group, it seems less ideal. In this analysis, they seemed to base their treatment decisions also on other factors than scientific evidences. This was also reflected in the comparison of the first, intuitive recommendation with the second, evidence-based decision. While medical graduate students constituted the most skeptical of all students about naturopathic treatment, physicians did not change their recommendation at all. In their free-text justification, physicians indicated patient wishes, placebo and the principle of doing no harm to be as important as scientific evidences (Figure 8). These features were summarized in a study as so-called contextual factors, comprising emotional reactions, behavioral inferences, optimizing the doctor patient relationship and difficulty with closure of the clinical encounter

(McBee et al., 2015). These factors have been found to be present and potentially impacting clinical reasoning in clinical practice. Generally, such features are considered an integral part of clinical decision-making, however, are often not fully evident to the physicians (Norman, Young, & Brooks, 2007). Another study found that social factors such as employment status of the patient and treatment duration are also taken into account when treatment alternatives for prostate cancer are considered (Davis et al., 2017). Generally, patient wishes and shared decision-making are one of the most important aspects in clinical reasoning (Driever, Stiggelbout, & Brand, 2020).

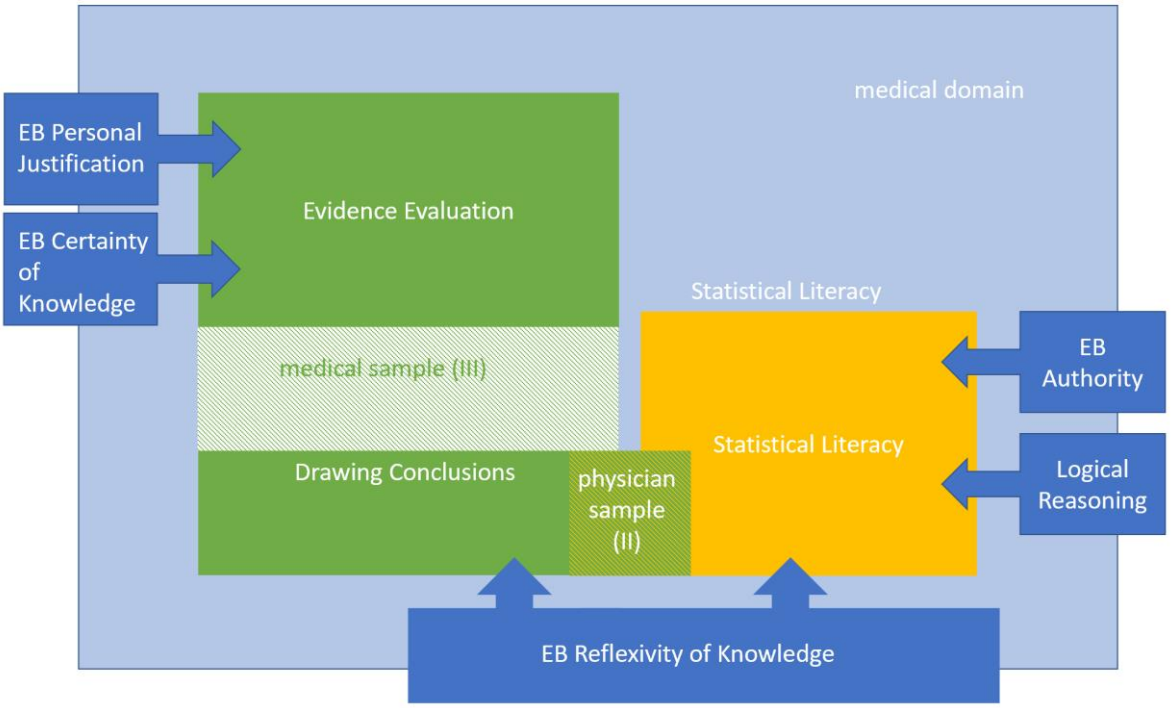
Taken together, this duality of scientific evidence and individual experiences (physician's expertise or patient specific factors) is reflected in the modern definition of evidence-based medicine (Sackett et al., 1997). This should be taken into account when analyzing drawing conclusions in physicians, especially in comparison to medical students, who are just developing these reasoning skills.

6.5.6. Model building with logical reasoning and epistemological beliefs

Based on this analysis in the medical domain, SL and SRA can be regarded as separate activities that can be learned, performed, fostered and applied separately (Figure 9). They are shaped by different epistemological beliefs, which change over the course of study and vary across domains. Statistical literacy is partly determined by logical reasoning, based on intellect and mathematical understanding. Evidence evaluation and drawing conclusions are shaped by different epistemological beliefs. In the medical sample, the correlation of EE and DC can be explained by the structure of the instrument, suggesting that the medical context of the decision scenario shaped this association, since it was not present in the student sample. The way of drawing conclusions in physicians differs from that of medical students.

Figure 9

Final Model



6.6. Strengths and Limitations

This dissertation introduced a new test instrument to measure SL and SRA skills applied in two different study samples. For the student sample, it additionally provided an in-depth measurement of covariables, such as logical reasoning, epistemological beliefs, and willingness to learn. For the physician sample, extensive analysis of demographic variables was included. This is a new approach to the measurement of these skills in the medical domain.

Due to the broad spectrum of item difficulty in SL, no ceiling effect - as observed in some other instruments - was present. This setup, however, let Cronbach's alpha fall below the .70 aspiration level in some scales in the student sample, which is common for short tests due to the mix of different test items (Cokely et al., 2012). The medical domain, in which the test and the scenario are designed, was not considered to impede test validity especially for students of the economic or social sciences domains. Furthermore, a study showed that "domain framing (…) did not necessarily differentially affect test performance or comprehension. This finding indicates that various domain-specific items (…) can provide a reasonable basis for the assessment of general statistical numeracy skills that will have predictive power across diverse domains" (Cokely et al., 2012, p. 26). In addition, Levy et al. (2014) found that health numeracy is strongly correlated with general numeracy.

The tool proved to be suitable to be applied in different study domains. It offers another assessment option apart from exam grades (Burkley & Burkley, 2009) or study-specific measures (Bachiochi et al., 2011) and allows to differentiate between well scoring students and those capable of demonstrating skills.

Regarding the measurement of SRA, there are two important points to consider. First, since the test instrument applied in this dissertation focuses only on two epistemological activities, i.e. evidence evaluation and drawing conclusion, and although it is combine with the assessment of SL and as such providing a broader perspective, it could be considered as too narrow to allow for new conceptualization of scientific reasoning (Adams & Wieman, 2015).

Second, the measurement of SRA skills with a scenario-based approach seems feasible and expedient. Therefore, future studies should seek to investigate if the domain context of the scenarios affects the measurement outcome of participants from different domains.

Concerning both, the assessment of SL and SRA, the data set used was cross-sectional with students from one university. Future studies should employ the tool in a longitudinal design (pre and post measurement) across several universities, to allow for the assessment of intra-personal development across institutions and curricula.

In the physician study group, the additional assessment of demographic variables was an asset It allowed an evaluation of how, when and where scientific skills were acquired and an analysis of the relevant factors. Although the study population was rather small, cell size was sufficient, and all tests were robust. No overrepresentation of physicians working in research or academia was observed. Although the test instrument has been validated and worked well with students, the decision scenario was not ideal to measure decision making abilities for physicians as their DC is mainly driven by patient wishes and other influencing factors.

7. Conclusion

Statistical literacy and SRA skills are essential aptitudes in a knowledge-society. They are especially important in medicine, a field in which research is the driving force of innovation and mere knowledge cannot be regarded as sufficient to ensure evidence-based and individualized patient care. Unfortunately, a collective deficit in SL as well as in SRA has been observed among physicians (Lipkus et al., 2001; Sedlmeier & Gigerenzer, 2001; Anderson et al., 2014).

This dissertation aimed at assessing SL and SRA in medical students and physicians in order to take stoke of the current situation, analyze these concepts in the medical domain in order to ultimately gain new information on how to foster these skills in medical education. Therefore, a new measurement tool was created, combining the assessment of SL and SRA skills and applied to medical students and physicians. The results have been presented in the three different parts of this dissertation.

The first part provided new insights in the distinct interplay of study domain and study progress on SL and SRA skills. Medical students receive better SL-Scores than social sciences students and comparable scores to economic students. A more advanced study phase alone did not constitute a sign for advanced skills. This could indicate that SL and SRA skills are not automatically developed over the course of higher education. Rather, they could depend on domain-specific characteristics, e.g. science curricula, appreciation of these skills by students and epistemological beliefs.

In the second part, this measurement tool was applied to physicians and extended with demographic variables. It showed that having worked in research and published papers were key factors in the development of SL and having completed a research projects for SRA. Since most participants indicated to have acquired scientific skills outside of formal medical education in an autodidactic manner, the active involvement in research within formal medical education constitutes a central element in fostering these skills.

The third part of this thesis integrated the two samples and allowed for a domain specific comparison of medical students and physicians. These skills are very similar in physicians and medical students of all study phases. Drawing conclusions in physicians follows the modern definition of evidence-based medicine (Sackett et al., 1997) and has thus to be analyzed differently.

Taken together, this thesis questioned the role of SL for SRA in the medical context and found that SL and SRA are separate skills which can be fostered individually. However, both skills are influenced by epistemological beliefs and logical reasoning.

In a lifelong learning setting, it seems important for the development of SL and SRA that medical students are involved in the process of research and writing papers already during their doctoral thesis. Furthermore, formal training of research and statistical skills should be continued during residency. Further research should examine the impact of science curricula in medical studies not only on SL and SRA skills, but also on epistemological beliefs.

References

Adams, W. K., & Wieman, C. E. (2015). Analyzing the many skills involved in solving complex phycics problems. *American Journal of Physics, 83*, 459-467. doi: 10.1119/1.4913923

Altman, D. G., & Bland, J. M. (1991). Improving doctors' understanding of statistics. *Journal of the Royal Statistical Society, Series A* (154), 223-67. doi:10.2307/2983040

Anderson, B. L., Gigerenzer, G., Parker, S., & Schulkin, J. (2014). Statistical literacy in obstetricians and gynecologists. *Journal of Healthcare Quality, 36*(1), 5-17. doi:10.1111/j.1945-1474.2011.00194.x

Anderson, B. L., Williams, S., & Schulkin, J. (2013). Statistical literacy of obstetrics-gynecology residents. *Journal of Graduate Medical Education, 5*(2), 272-275. doi:10.4300/JGME-D-12-00161.1

Ärzteblatt. (2017). Neue Empfehlungen zum Einsatz von Weißdornpräparaten bei Herzinsuffizienz. *Vermischtes.* Retrieved from https://www.aerzteblatt.de/fachgebiete/kardiologie/news?nid=72259

Bachiochi, P., Everton, W., Evans, M., Fugere, M., Escoto, C., Letterman, M., & Leszczynski, J. (2011). Using Empirical Article Analysis to Assess Research Methods Courses. *Teaching of Psychology, 38*(1), 5-9. doi:10.1177/0098628310387787

Bell, D. (1976). The coming of the post-industrial society. *The Educational Forum, 40*(4), 574-579). doi:10.1080/00131727609336501

Ben-Zvi, D., & Garfield, J. (2004). Statistical literacy, reasoning, and thinking: Goals, definitions, and challenges. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 3-25). Dordrecht, The Netherlands.

Broers, N. J. (2006). *Learning goals: The primacy of statistical knowledge.* Paper presented at the 7th Annual Meeting of ICOTS-/, Salvador, Brazil.

Budgett, S., & Rose, D. (2017). Developing statistical literacy in the final school year. *Statistics Education Research Journal, 16*(1), 139-162. Retrieved from https://iase-web.org/documents/SERJ/SERJ16(1)_Budgett.pdf

Burkley, E., & Burkley, M. (2009). Mythbusters: A Tool for Teaching Research Methods in Psychology. *Teaching of Psychology, 36*(3), 179-184. doi:10.1080/00986280902739586

Callingham, R., & Watson, J. M. (2017). The development of statistical literacy at school. *Statistics Education Research Journal, 16*(1), 181-201. Retrieved from http://iase-web.org/documents/SERJ/SERJ16(1)_Callingham.pdf

Carey, S., & Smith, C. (1993). On understanding the nature of scientific knowledge. *Educational Psychologist, 28*(3), 235-251. doi: 10.1207/s15326985ep2803_4

Castells, M. (1996). *The information age* (Vol. 98). Oxford Blackwell Publishers.

Cokely, E. T., Galesic, M., Schulz, E., Ghazal, S., & Garcia-Retamero, R. (2012). Measuring Risk Literacy: The Berlin Numeracy Test. *Judgment and Decision Making, 7*(1), 25-47. doi:10.1037/t45862-000

Covey, J. (2007). A meta-analysis of the effects of presenting treatment benefits in different formats. *Med Decis Making, 27*(5), 638-654. doi:10.1177/0272989X07306783

Csanadi, A., Eagan, B., Kollar, I., Shaffer, D. W., & Fischer, F. (2018). When coding-and-counting is not enough: using epistemic network analysis (ENA) to analyze verbal data in CSCL research. *International Journal of Computer-Supported Collaborative Learning, 13*(4), 419-438. doi:10.1007/s11412-018-9292-z

Davis, K., Bellini, P., Hagerman, C., Zinar, R., Leigh, D., Hoffman, R. et al. (2017). Physicians' perceptions of factors influencing the treatment decision-making process for men with low-risk prostate cancer. *Urology, 107*, 86-95. doi:10.1016/j.urology.2017.02.056

delMas, R., Ooms, A., Garfield, J., & Chance, B. (2006). Assessing students' statistical reasoning. In *Proceedings of the seventh international conference on teaching statistics* (July 2006). University of Minnesota.

Dietrich, H., Zhang, Y., Klopp, E., Brünken, R., Krause, U. M., Spinath, F. M. et al. (2015). Scientific Competencies in the Social Science. *Psychology Learning & Teaching, 14*(2), 115-130. doi:10.1177/1475725715592287

Dolan, J. G., Cherkasky, O. A., Li, Q., Chin, N., & Veazie, P. J. (2016). Should health numeracy be assessed objectively or subjectively? *Medical Decision Making, 36*(7), 868-875. doi: 10.1177/0272989X15584332

Driever, E. M., Stiggelbout, A. M., & Brand, P. L. P. (2020). Shared Decision Making: Physicians' preferred role, usual role and their perception of its key components. *Patient education and counseling, 103*(1), 77-82. doi:10.1016/j.pec.2020.02.019

Dunbar, K. (1995). How scientists really reason: Scientific reasoning in real-world laboratories. In R. J. Sternberg & J. Davidson (Eds.), *Mechanisms of insight*, (pp. 365-395). Cambridge, MA: MIT press.

Engelmann, K., Neuhaus, B. J., & Fischer, F. (2016). Fostering scientific reasoning in education–meta-analytic evidence from intervention studies. *Educational research and evaluation, 22*(5-6), 333-349. doi:10.1080/13803611.2016.1240089

Epstein, N., Huber, J., Gartmeier, M., Berberat, P., Reimer, M., & Fischer, M. (2018). Investigating on the acquisition of scientific competencies during medical studies and the medical doctoral thesis *Journal of Medical Education, 35*(2). doi: 10.3205/zma001167

Fagerlin, A., Zikmund-Fisher, B., Ubel, P., Jankovic, A., Derry, H., & Smith, D. (2007). Measuring Numeracy without a Math Test: Development of the Subjective Numeracy Scale. *Medical Decision Making, 27*, 672-680. doi:10.1177/0272989X07304449

Fischer, F., Kollar, I., Ufer, S., Sodian, B., Hussmann, H., Pekrun, R. et al. (2014). Scientific Reasoning and Argumentation: Advancing an Interdisciplinary Research Agenda in Education. *Frontline Learning Research*, 28-45. doi:10.14786/flr.v2i2.96

Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R. , Perry, M., & Scheaffer, R. (2005). *Guidelines for Assessment and Instruction in Statistics Education (GAISE) report: A Pre-K-12 curriculum framework.* American Statistical Association: Alexandria.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives, 19*, 25-42. doi:10.1257/089533005775196732

Gaissmaier, W., & Gigerenzer, G. (2008). Statistical illiteracy undermindes informed shared decision making. *Zeitschrift für Evidenz Fortbildung und Qualität im Gesundheitswesen, 8*(2), 53-96. doi:10.1016/j.zefq.2008.08.013

Gal, I. (2002). Adults´ statistical literacy: Meanings, components, responsibilities. *International statistical review, 70*(1), 1-51. Retrieved from http://iase-web.org/Publications.php?o=Int_Stat_Review

Garfield, J. (1995). How students learn statistics. *International statistical review, 63*(1), 25-34. doi:10.2307/1403775

Garfield, J. (2002). The challenge of developing statistical reasoning. *Journal of Statistics Education, 10*(3). doi:10.1080/10691898.2002.11910676

Ghanem, C., Kollar, I., Fischer, F., Lawson, T. R., & Pankofer, S. (2016). How do social work novices and experts solve professional problems? A micro-analysis of epistemic activities and the use of evidence. *European Journal of Social Work.* doi: 10.1080/13691457.2016.1255931

Ghazal, S., Cokely, E. T., & Garcia-Retamero, R. (2014). Predicting biases in very highly educated samples: Numeracy and metacognition. Judgment and Decision Making. *Judgment and Decision Making, 9*(1), 15-34

Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., & Woloshin, S. (2008). Helping Doctors and Patients Make Sense of Health Statistics *Psychological Science in the Public Interest, 8*(2), 53-96.

Gigerenzer, G., & Wegwarth, O. (2008). Risikoabschätzung in der Medizin am Beispiel der Krebsfrüherkennung. *Zeitschrift für Evidenz, Fortbildung und Qualität im Gesundheitswesen, 102*(9), 513-519.

Gormally, C., Brickman, P., & Lutz, M. (2012). Developing a test of scientific literacy skills (TOSLS): Measuring undergraduates´ evaluation of scienctific information and arguments. *Cell Biology Education, 11*, 364-377. doi:10.1187/cbe.12-03-0026

Griffith, J. D., Adams, L. T., Gu, L. L., Hart, C. L., & Nichols-Whitehead, P. (2012). Students´ attitudes towards statistics across the disciplines: a mixed-methods approach. *Statistics Education Research Journal, 11*(2), 45-56. doi:10.1037/t08417-000

Guilford, J. P. (1988). Some changes in the structure-of-intellect model. *Educational and Psychological Measurement, 48*(1), 1-4. doi:10.1177/001316448804800102

Hanoch, Y., Miron-Shatz, T., Cole, H., Himmelstein, M., & Federman, A. D. (2010). Choice, numeracy and physician-in-training performance: The case of Medicare part D. *Health Psychology, 29*, 454-459. doi:10.1037/a0019881

Harden, M., Grant, J., Buckley, G., Hart, R. (1999). BEME guide no. 1: best evidence medical education. *Medical teacher, 21*(6), 553-562. doi:10.1080/01421599978960

Hess, R., Visschers, V. H. M., Siegrist, M., & Keller, C. (2011). How do people perceive graphical risk communication? The role of subjective numeracy. *Journal of Risk Research, 14*(1), 47-61. doi:10.1080/13669877.2010.488745

Hetmanek, A., Engelmann, K., Opitz, A., & Fischer, F. (2018). Beyond intelligence and domain knowledge: Scientific reasoning and argumentation as a set of cross-domain skills. In F. Fischer, C. A. Chinn, K. Engelmann, & J. Osborne (Eds.), *Scientific reasoning and argumentation: The roles of domain-specific and domain-general knowledge* (pp. 203-226). New York: Routledge.

Hofer, B. K., & Pintrich, P. R. (2012). *Personal epistemology: The psychology of beliefs about knowledge and knowing*: New York: Routledge.

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence: An essay on the construction of formal operational structure*. London: Routledge & Kegan Pau.

Johnson, T. V., Abbasi, A., Schoenberg, E. D., Kellum, R., Speake, L. D., Spiker, C. et al. (2014). Numeracy among trainees: are we preparing physicians for evidence-based medicine? *Journal of surgical education, 71*(2), 211-215. doi:10.1016/j.jsurg.2013.07.013

Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive science, 12*(1), 1-48. doi:10.1016/0364-0213(88)90007-9

Klopp, E., & Stark, R. (2016). Persönliche Epistemologien - Elemente wissenschaftlicher Kompetenz. In A. K. Mayer & T. Rosman (Eds.), *Denken über Wissen und Wissenschaft. Epistemologische Überzeugungen als Gegenstand psychologischer Forschung* (pp. 39-69). Lengerich: Pabst Science Publishers.

Koslowski, B. (1996). *Theory and evidence: The development of scientific reasoning*. Cambridge, MA: MIT Press/ Bradford Books.

Kruglanski, A. W., & Gigerenzer, G. (2011). Intuitive and deliberate judgments are based on common principles. *Psychological Review, 118*(1), 97-109. doi:10.1037/a0020762

Laopaiboon, M., Lumbiganon, P., & Walter, S. D. (1997). Doctors' statistical literacy: a survey at Srinagarind Hospital, Khon Kaen University. *Journal of the Medical Association of Thailand= Chotmaihet thangphaet, 80*(2), 130-137.

Lederman, N. G. (2007). Nature of science: Past, present, and future. *Handbook of research on science education, 2*, 831-879.

Leira, E., Granner, M., Torner, J., Callison, R., & Adams, H. (2008). Education research: the challenge of incorporating formal research methodology training in a neurology residency. *Neurology, 70*(20), e79-e84. doi:10.1016/S2173-5808(11)70058-X

Lenzer, B., Ghanem, C., Weidenbusch, M., Fischer, M. R., & Zottmann, J. (2017). *Scientific Reasoning in Medical Education: A Novel Approach for the Analysis of Epistemic Activities in Clinical Case Discussions.* Paper presented at the Conference of the Association for Medical Education in Europe (AMEE), Helsinki, Finland.

Levy, H., Ubel, P. A., Dillard, A. J., Weir, D. R., & Fagerlin, A. (2014). Health Numeracy: The Importance of Domain in Assessing Numeracy. *Medical Decision Making, 34*, 107-115. doi:10.1177/0272989X13493144

Lin, D., Wei, X., & Mollohan, K. (2016). Does higher education improve student scientific reasoning skills? *International Journal of Science and Mathematics Education, 14*(4), 619-634. doi:10.1007/s10763-014-9597-y

Lipkus, I. M., Samsa, G., & Rimer, B. K. (2001). General performance on a numeracy scale among highly educated samples. *Medical Decision Making, 21*(1), 37-44. doi:10.1177/0272989X0102100105

Lloyd, S. A., & Robertson, C. L. (2012). Screencast tutorials enhance student learning of statistics. *Teaching of Psychology, 39*(1), 67-71. doi:10.1177/0098628311430640

Martin, N., Hughes, J., & Fugelsang, J. (2017). The roles of experience, gender, and individual differences in statistical reasoning. *Statistics Education Research Journal, 16*(2). Retrieved from http://iase-web.org/Publications.php?p=SERJ

McBee, E., Ratcliffe, T., Picho, K., Artino, A. R., Schuwirth, L., Kelly, W., . . . Durning, S. J. (2015). Consequences of contextual factors on clinical reasoning in resident physicians. *Advances in Health Sciences Education, 20*(5), 1225-1236. doi:10.1007/s10459-015-9597-x

McKenzie, J. D. J. (2004). *Conveying the core concepts.* Paper presented at the Joint Statistical Meetings, Toronto. Retrieved from www.statlit.org/pdf/2004McKenzieASA.pdf

Meissner, T. (2017, 15. März 2017). Weißdorn: Effekt auf das Endothel im Fokus. *ÄrzteZeitung.* Retrieved from https://www.aerztezeitung.de/medizin/krankheiten/herzkreislauf/herzinsuffizienz/article/931560/weissdorn-effekt-endothel-fokus.html?sh=1&h=562368279

Msaouel, P., Kappos, T., Tasoulis, A., Apostolopoulos, A. P., Lekkas, I., Tripodaki, E.-S., & Keramaris, N. C. (2014). Assessment of cognitive biases and biostatistics knowledge of medical residents: a multicenter, cross-sectional questionnaire study. *Medical Education Online, 19*, 23646-23646. doi:10.3402/meo.v19.23646

Norman, G., Young, M., & Brooks, L. (2007). Non-analytical models of clinical reasoning: the role of experience. *Medical education, 41*(12), 1140-1145. doi:10.1111/j.1365-2923.2007.02914.x

Okamoto, M., Kyutoku, Y., Sawada, M., Clowney, L., Watanabe, E., Dan, I., & Kawamoto, K. (2012). Health numeracy in Japan: measures of basic numeracy account for framing bias in a highly numerate population. *BMC Med Inform Decis Mak, 12(*104). doi:10.1186/1472-6947-12-104

Opitz, A., Heene, M., & Fischer, F. (2017). Measuring scientific reasoning - a review of test instruments. *Educational research and evaluation, 23*(3-4), 78-101. doi: 10.1080/13803611.2017.1338586

Osborne, J. (2010). Arguing to learn in science: The role of collaborative, critical discourse. *Science, 328*(5977), 463-466. doi:10.1126/science.1183944

Peters, E. (2012). Beyond Comprehension: The Role of Numeracy in Judgments and Decisions. *Current Directions in Psychological Science, 21*(1), 31-35. doi: 10.1177/0963721411429960

Porsch, T., & Bromme, R. (2011). Effects of epistemological sensitization on source choices. *Instructional Science, 39*(6), 805-819. doi:10.1007/s11251-010-9155-0

Reyna, V. F., Nelson, W. L., Han, P. K., & Dieckmann, N. F. (2009). How numeracy influences risk comprehension and medical decision making. *Psychol Bull, 135*(6), 943-973. doi:10.1037/a0017327

Rieckmann, M. (2012). Future-oriented higher education: Which key competencies should be fostered through university teaching and learning? *Futures, 44*(2), 127-135. doi:10.1016/j.futures.2011.09.005

Riegelman, R. K., & Hovland, K. (2012). Scientific Thinking and Integrative Reasoning Skills (STIRS): Essential outcomes for medical education and for liberal education. *Peer Review, 14*(4), 10.

Rudolph, J. L., & Horibe, S. (2016). What do we mean by science education for civic engagement? *Journal of Research in Science Teaching, 53*(6), 805-820. doi: 10.1002/tea.21303

Rychen, D. S., & Salganik, L. H. (2003). *Key competencies for a successful life and well-functioning society.* Göttingen: Hogrefe Publishing.

Sackett, D. L., Rosenberg, W. M. C., Gray, J. A. M., Haynes, R. B., & Richardson, W. S. (1997). Was ist Evidenz-basierte Medizin und was nicht? *MMW Munchener Medizinische Wochenschrift, 139*(44), 28-29. doi:10.1055/b-0036-140841

Schmidt, R. L., Chute, D. J., Colbert-Getz, J. M., Firpo-Betancourt, A., James, D. S., Karp, J. K. et al. (2017). Statistical Literacy Among Academic Pathologists: A Survey Study to Gauge Knowledge of Frequently Used Statistical Tests Among Trainees and Faculty. *Arch Pathol Lab Med, 141*(2), 279-287. doi:10.5858/arpa.2016-0200-OA

Schneewind, K. A., & Graf, J. (1998). *Der 16 Persönlichkeits-Faktoren-Test. Revidierte Version (16 PF-R)*. Bern: Huber.

Schommer-Aikins, M. (2004). Explaining the epistemological belief system: Introducing the embedded systemic model and coordinated research approach. *Educational Psychologist, 39*(19-29). doi:10.1207/s15326985ep3901_3

Schommer, M. (1990). Effects of beliefs about the nature of knowledge on comprehension. *Journal of Educational Psychology, 82*, 498-504. doi:10.1037/0022-0663.82.3.498

Schuler, H., & Fintrup, A. (2002). Das Leistungsmotivationsinventar (Achievement Motivation Inventory). *Wirtschaftspsychologie, 9*(2), 78-82. doi:10.34156/9783791035123-546

Schuler, H., & Prochaska, M. (2003). Leistungsmotivationsinventar (LMI). In J. Erpenbeck & L. von Rosenstiel (Eds.), *Handbuch Kompetenzmessung. Erkennen, verstehen und bewerten von Kompetenzen in der betrieblichen, pädagogischen und psychologischen Praxis*. Stuttgart: Schäffer-Poeschel.

Schwartz, L. M., Woloshin, S., Black, W. C., & Welch, H. G. (1997). The role of numeracy in understanding the benefit of screening mammography. *Ann Intern Med, 127*(11), 966-972. doi:10.7326/0003-4819-127-11-199712010-00003

Schwartz, L. M., Woloshin, S., & Welch, H. G. (2005). Can patients interpret health information? An assessment of the medical data interpretation test. *Med Decis Making, 25*(3), 290-300. doi:10.1177/0272989X05276860

Sedlmeier, P., & Gigerenzer, G. (2001). Teaching Bayesian Reasoning in Less Than Two Hours. *Journal of Experimental Psychology: General, 130*(3), 380-400. doi: 10.1037/0096-3445.130.3.380

Shavelson, R. J., & Huang, L. (2003). Responding responsibly. *Change: The magazine of higher learning, 35*(1), 10-19. doi:10.1080/00091380309604739

Shield, M. (1999). Statistical literacy: thinking critically about statistics. *Of Significance, 1*(1), 15-20.

Shield, M. (2017). GAISE 2016 Report promotes statistical literacy. *Statistics Education Research Journal, 16*(1), 50-54.

Simon, D. (2016). Chronische Herzinsuffizienz. *Gesundheit heute.* Retrieved from http://www.apotheken.de/gesundheit-heute-news/article/chronische-herzinsuffizienz/

Snee, R. D. (1990). Statistical thinking and its contribution to total quality. *The American Statistician, 44*(2), 116-121. doi:10.2307/2684144

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics* (4th ed. ed.). Needham Heights, MA: Allyn & Bacon.

Timmerman, B. E. C., Strickland, D. C., Johnson, R. L., & Payne, J. R. (2011). Development of a ‘universal’rubric for assessing undergraduates' scientific reasoning skills using

scientific writing. *Assessment & Evaluation in Higher Education, 36*(5), 509-547. doi: 10.1080/02602930903540991

Tognolini, J. (1996). *Rasch modelling: Advantages and limitations.* Paper presented at the Session notes National Meeting on Assessment and Reporting 25-26 November 1996.

Urhahne, D. (2006). Die Bedeutung domänenspezifischer epistemologischer Überzeugungen für Motivation, Selbstkonzept und Lernstrategien von Studierenden. [The importance of domain-specific epistemological beliefs for students' motivation, self-concept, and learning strategies]. *Zeitschrift für Pädagogische Psychologie, 20*(3), 189-198. doi:10.1024/1010-0652.20.3.189

Välimaa, J., & Hoffman, D. (2008). Knowledge society discourse and higher education. *Higher Education, 56*(3), 265-285. doi:10.1007/s10734-008-9123-7

Veilleux, J. C., & Chapman, K. M. (2017). Development of a Research Methods and Statistics Concept Inventory. *Teaching Of Psychology, 44*(3), 203-211.

Walker, H. M. (1951). Statistical literacy on the social science. *The American Statistician, 5*(1), 6-12. doi:10.1080/00031305.1951.10481912

Wallman, K. K. (1993). Enhancing statistical literacy: enriching our society. *Journal of the American Statistical Association, 88*(421), 1-8. doi:10.2307/2290686

Watson, J. M. (1997). Assessing Statistical Thinking Using the Media. In I. Gal & J. B. Garfield (Eds.), *The Assessment Challenge in Statistics Education* (pp. 107-121). Amsterdam: IOS Press and The International Statistical Institute

Watson, J. M., & Callingham, R. (2003). Statistical literacy: A complex hierarchical construct. *Statistics Education Research Journal, 2*(2), 3-46.

White, B., Stains, M., Escriu-Sune, M., Medaglia, E., Rostamnjad, L., Chinn, C., & Sevian, H. (2011). A Novel Instrument for Assessing Students' Critical Thinking Abilities. *Journal of College Science Teaching, 40*(5). doi:10.1187/cbe.11-09-0086

Windish, D. M., Huot, S. J., & Green, M. L. (2007). Medicine residents' understanding of the biostatistics and results in the medical literature. *JAMA, 298*(9), 1010-1022. doi:10.1001/jama.298.9.1010

Wissenschaftsrat. (2014). Empfehlungen zur Weiterentwicklung des Medizinstudiums in Deutschland auf Grundlage einer Bestandsaufnahme der humanmedizinischen Modellstudiengänge, Dresden: Wissenschaftsrat.

Zieffler, A., Garfield, J., Alt, S., Dupuis, D., Holleque, K., & Chang, B. (2008). What does research suggest about the teaching and learning of introductory statistics at the collewge level - A review of the literature. *Journal of Statistics Education, 16*(2). doi: 10.1080/10691898.2008.11889566

Zimmerman, C. (2000). The development of scientific reasoning skills. *Developmental Review, 20*, 99-149. doi:10.1006/drev.1999.0497

Zimmerman, C. (2007). The development of scientific thinking skills in elementary and middle school. *Developmental Review, 27*, 172-223. doi:10.1016/j.dr.2006.12.001

Acknowledgement

Firstly, I would like to express my sincere gratitude to my supervisor Dr. Markus Berndt for his continuous support, for his guidance, his insightful comments and encouragements. I could not have imagined a better colleague in the project ForschenLernen, nor a better supervisor for my doctoral thesis.

Furthermore, I want to thank my co-authors Jan Zottmann, for his valuable contributions and Prof. Dr. Maximilian Sailer, for the continuous support with the data analyses and Prof. Dr. med. Martin R. Fischer for providing the research environment and getting me involved in the research activities in medical school. He constitutes the link between research and education, so that this work may actually contribute to medical curriculum development in the future.