# Core ideas for teaching hypothesis testing – Structure, concepts and validation

Henrik Ossadnik and Jürgen Roth
RPTU University of Kaiserslautern-Landau, Germany
h.ossadnik@rptu.de

*Hypothesis testing is often perceived as an inaccessible and difficult to teach topic in school education, partly due to weak curricular connections across lower and upper secondary levels. The complexity of key statistical concepts further contributes to this challenge. Additionally, instruction often emphasizes computational procedures over conceptual understanding and reasoning. To address these issues, we propose a spiral curriculum that utilizes core ideas as conceptual anchors to support early informal understanding and facilitate the transition to more formal understanding of hypothesis testing. Four preliminary core ideas were theoretically derived and validated through an expert survey. Initial findings confirm the relevance of the core ideas while highlighting the need to modify the core ideas' content and to adjust the relationships among them. This underscores the importance of further refinement of the theoretical framework prior to classroom implementation and lays the foundation for developing learning environments within a design-based research.*

## PROBLEM OUTLINE AND OBJECTIVES

The ability to analyze, interpret, and responsibly use data to make decisions in everyday life is becoming increasingly important. Gal's (2002) framework of "statistical literacy" captures these competencies. To anchor statistical literacy at the beginning of the 21st century in school, international frameworks also highlighted the importance of statistics education with data analysis and probability as central curriculum strands (see NCTM standards in NCTM 2001) and framed statistical literacy as a central learning goal (see GAISE guidelines, Franklin et al., 2005; Franklin & Bargagliotti, 2020). In Germany, the KMK standards (2012; 2022a and 2022b, each with the first version from 2004) have established the "Leitidee Daten und Zufall" [Guiding principle: data and chance] across the school curriculum. Although it places less emphasis on statistics than the international frameworks, inferential reasoning is still explicitly addressed, for instance, by interpreting hypothesis tests and justifying uncertainty of conclusions (KMK, 2012; Arbeitskreis Stochastik, 2003; Biehler, Engel & Frischemeier, 2023). These reforms already responded to the growing societal importance of data and statistics (Burrill & Biehler, 2011; Biehler, Engel & Frischemeier, 2023). Furthermore, didactical contributions in stochastics education emphasize the need to foster reasoning, interpretation, and argumentation, as opposed to an exclusive focus on procedural skills (Krüger, Sill & Sikora, 2015; Arbeitskreis Stochastik, 2003). But still, many students often do not develop the intended competencies at the end of high school (Rolfes & Heinze, 2022) – for example, in inferential statistics. One possible explanation for this phenomenon is the conceptually demanding nature of statistical inference, which is difficult to conceptualize, understand, and teach (Borovcnik, 2019). Research shows that students and teachers struggle with core inferential notions (e.g. in understanding hypothesis testing) and that teachers share misconceptions with their students. Consequently, incorporating inference (especially before university) is a challenge and further raises questions about its coherent progression from lower into higher secondary education (Lugo-Armenta & Pino-Fan, 2021; Batanero, Burrill & Reading, 2011). Moreover, it appears that instruction of inferential statistics is more about computational procedures than interpretative reasoning. This restricts students' opportunities to develop a deeper understanding of core statistical ideas (Hauer-Typpelt, 2010; Rolfes & Heinze, 2022). In addition, researchers have highlighted that the complexity of certain highly abstract concepts, such as statistical distributions, sampling variability, sampling distributions, and significance are perceived by students as complex and counterintuitive (Ben-Zvi & Garfield, 2004; see also the studies in the same volume). For instance, to understand sampling distributions, one must first grasp the fundamental concepts related to samples, as well as their interrelationships with other concepts, including variability, distribution, and center (Ben-Zvi et al., 2015). These foundations should be introduced early, so that students already possess an understanding of these concepts by the time more formal ideas of sampling distributions are addressed (Garfield & Ben-Zvi, 2008). More broadly, statistical literacy develops over time and necessitates the continuous integration of previously acquired alongside new ones, rather than addressing them all at

In: Birk, L., Loth, G., Jotzo, L., Binder, K., & Frischemeier, D. (Eds.) (2026). *Statistics and Data Science Education in STEAM. Proceedings of the Satellite Conference of the International Association for Statistical Education (IASE), September/October 2025, Münster, Germany*. © ISI/IASE.

once (Watson, 2006). This ensures the development of a solid conceptual understanding and supports drawing appropriate conclusions. On the other hand, this implies that, if this integrative process begins only at the upper secondary level, many key ideas must be introduced concurrently and directly connected. This may result in a fragmented understanding and challenges with overall comprehension.

As an early and potential pathway, informal inferential reasoning (IIR) has been discussed to deepen learners' understanding of statistical processes and outcomes. Introducing IIR at an early stage and revisiting it over time provides opportunities to facilitate engagement with statistical concepts, rather than encountering them only through formal procedures later (Makar & Rubin, 2009). This aligns with the idea of a spiral curriculum approach.

Building on this, it is important to consider which specific elements of inferential statistics may benefit from an early, informal introduction. This research project focuses on hypothesis testing as a central and exemplary procedure in inferential statistics, which is both a demanding and often misinterpreted procedure. We examine which aspects of informal hypothesis testing can be addressed early to prepare students for formal reasoning. To pursue this aim, we formulate a set of preliminary core ideas as a working basis, to support a spiral learning approach, and to facilitate classroom implementation. On this basis, our central research question is: *Which core ideas are appropriate for fostering students' informal understanding of hypothesis testing and structuring instruction in this area*? To concretize the research question, we further ask: (1) How can these core ideas be theoretically substantiated? (2) How do experts evaluate them and what feedback do they give? and (3) What modifications are necessary based on this expert feedback?

## WORKING CONCEPT: CORE IDEAS

We adapted the concept of core ideas (see Leuders et al., 2011) to teach statistics in a spiral curriculum and conceptualize them as elements of understanding for instruction. Core ideas refer to recurring and sustainable concepts within a given subject area. In combination with content-related anchors of understanding (Roth & vom Hofe, 2023), they provide a basis for planning teaching-learning processes and thus serve as a framework for instructional design. Through their mediating role between the subject-specific core and the learners' conceptual level, core ideas play an important role in building and deepening understanding. By linking this subject-specific (retrospective) and learner-oriented (prospective) perspectives, core ideas help to clarify the meaning of statistical concepts and reveal interconnections. In this sense, core ideas have the potential to support informal statistical inference and provide a bridge to more formal inferential statistics.

## DERIVE AND VALIDATE CORE IDEAS FOR STRUCTURING HYPOTHESIS TESTING

To address the central research question, the project followed a two phase-approach: (1) a theoretical derivation, through which we developed a set of preliminary core ideas related to hypothesis testing, and (2) an expert validation, in which these core ideas were evaluated, refined where necessary and completed by additional ideas that might enhance the initial framework.

### *Theoretical derivation process*

We derived the preliminary set of core ideas in a multi-step process: (1) a literature review to identify the key concepts, (2) a retrospective classification from a professional, subject-specific perspective, and (3) a prospective meta-reflection of subject-specific concepts from the learner's point of view. This process linked the two perspectives, clarified the meaning of the concepts, and simultaneously allowed us to condense and conceptualize core ideas. The goal was to distill the core ideas into concepts suitable for integration into a spiral curriculum, with a chance of being incorporated into the actual teaching practice. This process resulted in four preliminary core ideas (see Figure 1), which will be presented later in combination with the results of the expert survey.

### *Expert survey validation*

We subsequently conducted an expert survey to validate the core ideas from different perspectives and to establish a consensus. The survey pursued four objectives, (1) to assess the relevance of the core ideas from a didactic and subject-specific perspective; and evaluate their potential to support an understanding of hypothesis testing, (2) to identify ambiguities and refine or expand individual core

ideas to support the development process, (3) to determine necessary modifications and, (4) to add further core ideas, if required. To ensure the broadest possible and practice-oriented evaluation, we involved diverse experts with different professional backgrounds, including statistics educators, applied statisticians, school teachers, and school-university intermediaries, such as delegated teachers. We subdivided the expert survey into two steps:

1. **Video-based questionnaire:** The online questionnaire comprised both closed and open response formats and was introduced with a short video-based explanatory input on the above-explained working concept of core ideas. Each of the four preliminary ideas was then briefly presented and conceptualized based on a video input. We chose this input form to enhance the visualization of interrelationships and prevent superficial reading. We invited the participants to critically evaluate the core ideas with respect to their relevance, content, differences from other core ideas, and possible additions. In an exploratory section, experts could suggest alternative or additional core ideas, elaborate on them, or highlight further aspects that should be considered. Finally, respondents could opt for an in-depth stimulated recall interview to elaborate on their written feedback, thus allowing us to gain a comprehensive understanding of their perspectives.
2. **In-depth stimulated recall interview:** The interviews were based on the participants' questionnaire responses, which served as stimuli. In the initial part of the interview, the experts freely shared their insights, incorporating the core concept within the hypothesis test framework. They explained their reasoning, added further aspects to the written material, and highlighted aspects they considered particularly important. The semi-structured format provided a clear structure but left enough space for further questions. To clarify ambiguities arising from text-based feedback, we prepared several specific in-depth questions. All interviews were recorded, fully transcribed, and edited for analysis.

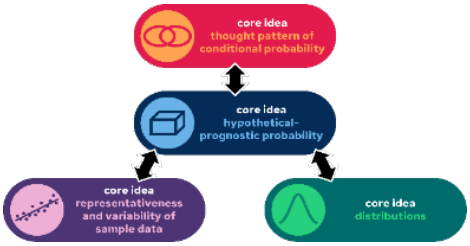


Figure 1. Preliminary core ideas for teaching hypothesis testing.

## ANALYSIS

The collected data were analyzed using qualitative content analysis (Kuckartz, 2018). The first author performed the coding with units of meaning that could be understood independently of the questionnaire. The entire questionnaire and the videos served as the context unit, as the videos were frequently referenced. We developed deductive-inductive coding. The deductive (*a priori*) categories were derived directly from the questionnaire structure, e.g., "Justification and evaluation of the relevance of the core ideas", "Justification of unrelated aspects", "Explanation of additional aspects", "Alternative title suggestions", and "Justification of alternative title suggestions". Although the specific questions focused intensely on these categories, overlapping responses were coded across categories where appropriate, and multiple coding was allowed. In addition, inductive (*a posteriori*) categories emerged based on the experts' feedback, e.g., "Vagueness in the presentation of the core idea", "Weighting of the core idea," or "Level of the core idea in the overall context", "Problems regarding aspect XY". The coding proceeded in multiple stages: first, we categorized the responses into the addressed units of meaning and assigned them to the deductive categories. Then, we carried out the inductive subcategorization of these statements by comparing them. The coding guide was developed based on the first seven questionnaires and has been continuously revised since then. Each category was defined by a preliminary coding rule and anchor examples.

We are currently coding the interviews using the adapted coding guide, which will be refined if necessary. The interview coding checks whether the categories coded in the questionnaire are also reflected in the interview statements and adds new core ideas, explanations, or suggestions for restructuring the overall concept, mentioned only in the interviews. Lastly, we will conduct an intercoder reliability check to verify the coding guide.

## SAMPLE OF EXPERTS SURVEYED

The sample consists of 12 experts: 11 completed the questionnaire, eight participated in a stimulated recall interview, and one person participated exclusively in the interview. In order to obtain a broad perspective, we selected a heterogeneous composition of the experts (the number of persons is given in brackets): As mentioned above, the sample included statistics educators (5), school-university intermediaries (2), school teachers (3), and professional mathematicians (3), including one statistician and one expert in financial mathematics. All participants identified themselves as experts in statistics and stochastics and reported regular use of hypothesis testing, at least once a quarter in their work. The participants were recruited from various universities in german-speaking countries through targeted invitations. They held at least a doctoral degree, and many of them were professors. Their professional experience ranged from 20 to 40 years. All in all, this confirms their suitability for this study.

This article presents the results of the questionnaire analysis and identifies areas where the conceptualization of individual core ideas and the overall framework may require revision. The presentation at the conference will summarize the results of the entire expert survey. The following first outlines each theoretically derived preliminary core idea separately, complements them with the expert feedback from the questionnaires, and discusses preliminary suggestions for potential modifications.

## CORE IDEA: HYPOTHETICAL-PROGNOSTIC PROBABILITY

The core idea of *hypothetical-prognostic probability* (cf. Riemer, 2019; Riemer, 2023) is positioned centrally (see Figure 1), as it serves as an intermediary between all other core ideas. The core idea stands for a concept of probability that also supports inferential statistical thinking – echoing previous calls to integrate probability explicitly in this role (Hauer-Typpelt, 2022; Burrill & Biehler, 2011). This conception perceives probabilities as mathematical models that are never "true" but always hypothetical. This modeling approach involves two perspectives: (1) prognostic: probabilities are mathematical models that predict the fluctuations in relative frequencies when we repeat the experiment multiple times, and (2) hypothetical: we can doubt models and modify them if the predictions do not reflect reality well when we repeat the experiment multiple times. This concept of probability establishes a link between probabilities and relative frequencies, and thus can be linked to ideas about inferential statistics at the high school level (Körner & Riemer, 2019). Through this understanding of the model-based nature of probabilities and the reflective process of questioning the adequacy of probability models, this core idea informally prepares students for the logic of hypothesis testing.

The consensus of the experts was that this is a core idea for hypothesis testing. Out of eleven responses analyzed, seven respondents selected "strongly agree", three selected "somewhat agree", and only one respondent expressed low agreement with "somewhat disagree" to rate the core idea. The experts emphasized that probabilities are only a model for predicting future relative frequencies and noted that this core idea is appropriate for use in schools.

At the same time, the experts also pointed out the need for modifications. They mentioned several times that this concept of probability is a fundamental basis for teaching stochastics in general and criticized its exclusive emphasis on hypothesis testing, which could be misleading. One expert noted: "Of course in hypothesis testing, but I would see it more as a general foundation." This leads us to conclude that we can distinguish between overarching, more general core ideas that lead to hypothesis testing and more specific core ideas of hypothesis testing. The current strong high school-centered focus on the binomial test, and thus the universal applicability of this core idea to other tests, was also critiqued by individual experts. Finally, a few responses referred to problems with the terminology. For the term "hypothetical-prognostic", one expert explained that this "pair [...] does not describe a separate concept of probability, but a property of probabilities in general". A stronger focus on the modelling nature of probabilities, explicitly in the title, may address this issue somewhat.

## CORE IDEA: REPRESENTATIVENESS AND VARIABILITY OF SAMPLE DATA

Hypothesis testing involves discussing the probability of an observed sample result about a hypothesized distribution. The core idea of *representativeness and variability of sample data* is crucial at this point because, by referring to an understanding of the generation and properties of appropriate samples, it targets a key aspect of inferential statistical thinking (Ben-Zvi et al., 2015; Watson, 2004; Garfield et al., 2015). The two interrelated concepts that are fundamental to this are: (1) the

representativeness of a sample with respect to the population, which affects how results can be generalized, and (2) the variability of the sample data. To protect against potential misunderstandings arising from overreliance on one concept over the other (Ben-Zvi et al., 2015), and to facilitate conceptual understanding, it is essential to know how both concepts complement and influence each other. This phenomenon is related to the development of a "hierarchical picture of samples". This picture integrates (1) the variation within a sample and (2) seeing a sample as part of a set of potential other samples (Lehrer, 2017; Saldanha & Thompson, 2014). Hypothesis testing deals with sample outcomes. Therefore, it is crucial to regard these results as a potential outcome and to support an important basis for informal hypothesis testing. A fundamental property of the law of large numbers can also be derived from this core idea: Larger samples tend to be less variable and generally more representative of the population (Bakker, 2004; Ben-Zvi et al., 2015).

The experts also rated the core idea of *representativeness and variability of sample data* as a core idea for hypothesis testing. Of the eleven responses received, six selected "strongly agree", four "somewhat agree", and only one expressed reservation with "somewhat disagree". They emphasized the importance of variability and the role of random samples as fundamental to inferential statistical reasoning and practice, noting that "the data are random, the result of the hypothesis test is also random, and therefore can be wrong".

Concurrently, the results of the questionnaire also suggest some modifications. The concept of representativeness was the central point of criticism for several respondents. First, experts pointed out that variability and representativeness are not contradictory concepts. Second, the experts noted that "representativeness [...] is a problematic term that is popular, but rarely precise". This points to the need for a more precise breakdown, coupled with the introduction of additional concepts. Simultaneously, it remains important to convey how a sample that is as representative as possible facilitates the appropriate interpretation of variability. Finally, the experts highlighted the concept of signal and suggested it should either stand as a core idea in its own right or be incorporated more explicitly: "Signal and noise [would] also be an important aspect, whose name better describes some aspects." This feedback indicates that the core idea could be strengthened by emphasizing the signal and noise concept and/or adjusting the title to clarify the connection to hypothesis testing.

## CORE IDEA: DISTRIBUTIONS

The concept of the core idea of *distributions* is based on the understanding of the ubiquity of variability, and statistics is concerned with how patterns in data can be understood in the midst of variability (Konold & Pollatsek, 2002; Garfield et al., 2008). Distributions play a crucial role in this search for explanations because they help to answer two questions: (1) How can the observed variability in the data be measured? And (2) How can the empirical variability be explained and predicted? Empirical distributions answer the first question by representing empirical data and allowing the variability to be interpreted using statistical ratios. In contrast, theoretical distributions are used to model and predict variability in reality. In this sense, hypothesis testing can be viewed as a negotiation process between empirical data (which can be described by empirical distributions) and theoretical distributions. These two perspectives converge here to answer the question of whether an observed deviation can be attributed to chance. Consequently, students need to understand the differences and interrelationships between these two main types of distributions. Closely related to this, students need to develop a global view of data as an aggregate, rather than just as a collection of individual values (Konold et al., 2015). This view allows us to (1) describe central features of distributions (e.g. shape, spread and center), which can be interpreted as model elements as they appear in the data, and (2) compare the data structure in one sample with that in others. Furthermore, central tendencies should always be considered in addition to dispersion (Watson, 2006; Konold & Pollatsek, 2002).

The majority also recognized the core idea of distributions as an integral part of hypothesis testing. Of the eleven experts, four "strongly agreed", six "somewhat agreed", and only one "somewhat disagreed". For this core idea, some experts emphasized that distributions are "at the center of hypothesis testing" and some emphasized that comparing empirical results and theoretical distributions is the "heart of hypothesis testing".

However, the experts also suggested several modifications, considering the title of the core idea to be potentially misleading and could lead to the erroneous suggestion that "it is more about concrete

distributions". However, the focus here is not on a specific distribution, but on comparing distributions, which is "much more important". For this reason, the experts suggested that this objective should already be made clear in the title. Given the potential centrality of this process, a revision of the weighting and hierarchy of the core ideas seems necessary. The dual function of "distribution" was also problematic, so that a more precise separation between model and reality at the linguistic level was requested. In addition, experts criticized the lack of concrete integration of sampling distribution and the link between this core idea and the previous one. The breadth of this core idea was also met with skepticism. As one respondent put it, "distributions are obviously important, but the explanations of how distributions are to be understood here make it very complex, and there are many references to other core ideas". This comment suggests that this core idea may be overloaded as presented: "The fact is that distributions play a central role, independent of hypothesis testing, as in probability theory or descriptive statistics". Considering these arguments, it may be necessary, as already mentioned before, to sharpen the focus on "core ideas that lead to hypothesis testing" and "core ideas of hypothesis testing" in general.

## CORE IDEA: THOUGHT PATTERN OF CONDITIONAL PROBABILITY

Despite the absence of assignable probabilities in classical hypothesis testing, the development of a conditional thought pattern and the associated understanding of "if ... then ..." structures is important for understanding hypothesis testing. In contrast to classical conditional probabilities, the conditional thought pattern weakens the condition to an assumption, for example, that a model is correct. Evaluating the plausibility of observed data under the assumption of such models can provide information that another model may be more accurate. However, it is impossible to make a statement about the probability of such valid models. This core idea is a form of p-value interpretation, which describes the probability that the sample result, or even more extreme values, occur under the assumption that $H_0$ holds, in conjunction with error probabilities.

The experts predominantly rated the core idea of the *thought pattern of conditional probability* as relevant to hypothesis testing. In this case, seven respondents "strongly agreed", one respondent "somewhat agreed", two respondents "somewhat disagreed", and one respondent "strongly disagreed". The experts justified the strong agreement by saying that the conditional structure discussed is very much in line with the logic of hypothesis testing and that the "thought pattern" presented this data evaluation under an assumed model hypothesis in an easily understandable way.

At the same time, we identified discernible impulses for modification. Several experts criticized the choice of terms in the title. The term "thought pattern" was then perceived as "artificial", and the established term "conditional probability" was described as "misleading". One response emphasized: "I can understand the content [...] very well, but the term 'conditional probability' does not capture it at all, it has completely different associations." The choice of the title should be reconsidered, with more emphasis on the conditional nature, but without using pre-existing and already well-established definitions. The underlying logic of the conclusion was endorsed on several occasions, although there were also calls for greater clarification of the logical structure behind a test. Some participants described the thought pattern as a variant of the mode of reasoning or the principle of contraposition or analogy to indirect proof: "You know: 1) If the model is valid, then certain data are probable. 2) These data did not occur. One concludes: The model is probably not valid." Finally, several experts suggested including the Bayesian perspective as a contrasting viewpoint to underscore the limitations of the frequentist interpretation. In their view, this could help to clarify the distinction between a mathematical condition and a model-based assumption.

## CONCLUSION

The initial findings from the questionnaire confirm the relevance of the four theoretically derived preliminary core ideas, while highlighting the need for modifications. These relate to the content of core ideas', the titles' wording, and the distinction between overarching core ideas that lead to hypothesis testing, and more specific core ideas within hypothesis testing. This raises the question about how core ideas develop and expand. Some experts also hinted at potential new core ideas, though not yet fully elaborated. The interviews will provide further clarification on the conceptualisation of the existing and the new potential core ideas, their conceptual acuity, and the conceptual structure and hierarchy that exists among them. In conclusion, these findings provide a foundation for a design-based

research project that aims at developing learning environments that foster students' conceptual understanding of the core ideas and support their understanding of hypothesis testing itself.

REFERENCES

Arbeitskreis Stochastik der GDM (2003). Empfehlungen zu Zielen und zur Gestaltung des Stochastikunterrichts [Recommendations on objectives and the structure of stochastics lessons]. *Stochastik in der Schule, 23*(3), 21–26.

Bakker, A. (2004). Reasoning about shape as a pattern in variability. *Statistics Education Research Journal, 3*(2), 64–83. https://doi.org/10.52041/serj.v3i2.552

Batanero, C., Burrill, G. & Reading, C. (2011). Overview: Challenges for teaching statistics in school mathematics and preparing mathematics teachers. In C. Batanero, G. Burrill & C. Reading (Eds.), *Teaching statistics in school mathematics – Challenges for teaching and teacher education. A Joint ICMI/IASE Study: The 18th ICMI Study* (pp. 407–418). Springer. https://doi.org/10.1007/978-94-007-1131-0

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. 3–15). Springer. https://doi.org/10.1007/1-4020-2278-6_1

Ben-Zvi, D., Bakker, A., & Makar, K. (2015). Learning to reason from samples. *Educational Studies in Mathematics, 88*(3), 291–303. https://doi.org/10.1007/s10649-015-9593-3

Biehler, R., Engel, J. & Frischemeier, D. (2023). Stochastik: Leitidee Daten und Zufall [Stochastics: Guiding principle Data and chance]. In R. Bruder, A. Büchter, H. Gasteiger, B. Schmidt-Thieme & H. G. Weigand (Eds.), *Handbuch der Mathematikdidaktik* (2nd edition, pp. 243–278). Springer. https://doi.org/10.1007/978-3-662-66604-3_8

Burrill, G., & Biehler, R. (2011). Fundamental Statistical Ideas in the School Curriculum and in Training Teachers. In C. Batanero, G. Burrill & C. Reading (Eds.), *Teaching statistics in school mathematics-Challenges for teaching and teacher education. A Joint ICMI/IASE Study: The 18th ICMI Study* (pp. 57–69). Springer. https://doi.org/10.1007/978-94-007-1131-0_10

Borovcnik, M. (2019). Informal inference – approaches towards statistical inference. In S. Budgett (Ed.), *Decision making based on data. Proceedings of the Satellite conference of the International Association for Statistical Education (IASE).* ISI/IASE.

Franklin, C., & Bargagliotti, A. (2020). Introducing GAISE II: A guideline for precollege statistics and data science education. *Harvard Data Science Review, 2*(4). https://doi.org/10.1162/99608f92.246107bb

Franklin, C., Kader, G., Mewborn, D. S., 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.

Gal, I. (2002). Adults' statistical literacy: Meanings, components, responsibilities. *International Statistical Review, 70*(1), 1–51. https://doi.org/10.2307/1403713

Garfield, J. B., Ben-Zvi, D., Chance, B., Medina, E., Roseth, C., & Zieffler, A. (2008). *Developing students' statistical reasoning*. Springer. https://doi.org/10.1007/978-1-4020-8383-9

Garfield, J., Le, L., Zieffler, A., & Ben-Zvi, D. (2015). Developing students' reasoning about samples and sampling variability as a path to expert statistical thinking. *Educational Studies in Mathematics*, *88*(3), 327–342. https://doi.org/10.1007/s10649-014-9541-7

Hauer-Typpelt, P. (2010). Angemessene Grundvorstellungen zu Wahrscheinlichkeit und Zufall entwickeln – Vorschläge für den Stochastikunterricht [Developing appropriate basic mental models of probability and chance – suggestions for teaching stochastics]. *Schriftenreihe zur Didaktik der Mathematik der Österreichischen Mathematischen Gesellschaft*, *43*.

Hauer-Typpelt, P. (2022). Stochastik in der Sekundarstufe 1 [Stochastics in lower secondary level]. *Schriftenreihe zur Didaktik der Mathematik der Österreichischen Mathematischen Gesellschaft*, *54*, 35–49.

Konold, C., & Pollatsek, A. (2002). Data analysis as the search for signals in noisy processes. *Journal for Research in Mathematics Education, 33*(4), 259–289. https://doi.org/10.2307/749741

Konold, C., Higgins, T., Russell, S. J., & Khalil, K. (2015). Data seen through different lenses. *Educational Studies in Mathematics, 88*(3), 305–325. https://doi.org/10.1007/s10649-013-9529-8

Körner, H., & Riemer, W. (2019). Beurteilende Statistik: ab Klasse 8! [Inferential statistics: from Year 8 onwards!]. *Der Mathematikunterricht, 65*(6), 4–10.

Krüger, K., Sill, H. D., & Sikora, C. (2015). *Didaktik der Stochastik in der Sekundarstufe I* [Teaching stochastics in lower secondary education]. Springer Spektrum. https://doi.org/10.1007/978-3-662-43355-3

Kuckartz, U. (2018). *Qualitative Inhaltsanalyse. Methoden, Praxis, Computerunterstützung* [Qualitative content analysis. Methods, practice, computer support] (4th ed.). Beltz Juventa.

KMK (Kultusministerkonferenz). (2012). *Bildungsstandards im Fach Mathematik für die Allgemeine Hochschulreife* [Educational standards in mathematics for the general higher education entrance qualification] (Beschluss der Kultusministerkonferenz vom 18.10.2012). KMK.

KMK (Kultusministerkonferenz). (2022a). *Bildungsstandards für das Fach Mathematik Primarbereich* [Educational standards for mathematics in primary education] (Beschluss der Kultusministerkonferenz vom 15.10.2004, i. d. F. vom 23.06.2022). KMK.

KMK (Kultusministerkonferenz). (2022b). *Bildungsstandards für das Fach Mathematik Erster Schulabschluss (ESA) und Mittlerer Schulabschluss (MSA)* [Educational standards for mathematics First school leaving certificate (ESA) and intermediate school leaving certificate (MSA)] (Beschluss der Kultusministerkonferenz vom 15.10.2004 und vom 04.12.2003, i. d. F. vom 23.06.2022). KMK.

Lehrer, R. (2017). Modeling signal-noise processes supports student construction of a hierarchical image of sample. *Statistics Education Research Journal*, *16*(2), 64–85. https://doi.org/10.52041/serj.v16i2.185

Leuders, T., Hußmann, S., Barzel, B., & Prediger, S. (2011). Das macht Sinn! Sinnstiftung mit Kontexten und Kernideen [That makes sense! Creating meaning with contexts and core ideas]. *Praxis der Mathematik in der Schule, 53*(37), 2–9.

Lugo-Armenta, J. G., & Pino-Fan, L. R. (2024). An approach to inferential reasoning levels on the Chi-square statistic. *Eurasia Journal of Mathematics, Science and Technology Education, 20*(1), Article em2388. https://doi.org/10.29333/ejmste/14119

Makar, K., & Rubin, A. (2009). A framework for thinking about informal statistical inference. *Statistics Education Research Journal, 8*(1), 82–105. https://doi.org/10.52041/serj.v8i1.457

National Council of Teachers of Mathematics (NCTM). (2001). NCTM Standards – Prinzipien und Standards für Schulmathematik: Datenanalyse und Wahrscheinlichkeit [NCTM Standards – Principles and Standards for School Mathematics: Data Analysis and Probability]. In M. Borovcnik, J. Engel & D. Wickmann (Eds.), *Anregungen zum Stochastikunterricht – Arbeitsbericht des AK Stochastik 1999/2000* (pp. 11–42). Franzbecker.

Riemer, W. (2019). Grundvorstellungen beurteilender Statistik: Wahrscheinlichkeit als bezweifelbares Modell [Basic mental models of inferential statistics: probability as a doubtful model]. *Der Mathematikunterricht, 65*(6), 11–22.

Riemer, W. (2023). *Statistik unterrichten: Eine handlungsorientierte Didaktik der Stochastik* [Teaching statistics: An action-oriented approach to teaching stochastics]. Klett Kallmeyer.

Rolfes, T., & Heinze, A. (2022). Vertiefte Allgemeinbildung als eine Zieldimension von Mathematikunterricht in der gymnasialen Oberstufe [In-depth general education as a target dimension of mathematics teaching in upper secondary school]. In T. Rolfes, S. Rach, S. Ufer & A. Heinze (Eds.), *Das Fach Mathematik in der gymnasialen Oberstufe* (pp. 19–46). Waxmann.

Roth, J., & vom Hofe, R. (2023). Verständnisvoll lernen – Grundvorstellungen vernetzen und Verständnisanker nutzen [Learning with understanding – linking basic mental models and using anchors of understanding]. *Mathematik lehren, 236*, 8–11.

Saldanha, L. A., & Thompson, P. W. (2014). Conceptual issues in understanding the inner logic of statistical inference: Insights from two teaching experiments. *The Journal of Mathematical Behavior*, *35*, 1–30. https://doi.org/10.1016/j.jmathb.2014.03.001

Watson, J. M. (2006). *Statistical literacy at school: Growth and goals*. Routledge. https://doi.org/10.4324/9780203053898

Watson, J. M. (2004). Developing Reasoning about Samples. In D. Ben-Zvi & J. Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning and Thinking* (pp. 277–294). Springer. https://doi.org/10.1007/1-4020-2278-6_12