MAPPING KNOWLEDGE FOR PROBABILITY AND STATISTICS APPLICATION: MATHEMATICAL AND NON-MATHEMATICAL

Sylvia Kuzmak
Rise Coaching and Consulting LLC, NJ, USA
skuzmak@uchicago.edu

A mapping (a cognitive analysis) of knowledge used in the effective application of probability and statistics in real world situations is presented, identifying both mathematical and non-mathematical knowledge involved. The fundamental role of domain-related knowledge and expertise is emphasized, with illustrations provided using applications involving social data. Use of the knowledge mapping to support improvement of educational practice is proposed, including its use as a basis to evaluate the completeness of current curriculum and instruction in applied probability and statistics, to guide the design of more complete and effective curriculum and instruction, to guide curriculum design across university departments, and to evaluate curriculum and instruction at all grade levels to ensure misapplication of probability and statistics is not taught.

INTRODUCTION

The power and value of probability and statistics lies in its appropriate application. Gigerenzer et al. (1989, pp. xiii-xiv) have noted the close relationship between probability theory and its applications, noting that advances in probability theory went along with expansion of the range of domains to which it was applied, which began with gambling problems in the mid-seventeenth century, went on “to jurisprudence, data analysis, inductive inference, and insurance” in the eighteenth, “and from there to sociology, to physics, to biology, and to psychology” in the nineteenth, “and on to agronomy, polling, medical testing, baseball …” and numerous others in the twentieth. The inspiration and payoff of probability and statistics has and continues to reside in its appropriate application in a range of domains.

While today we categorize probability and statistics as a branch of mathematics, and teach “the mathematics” of probability and statistics, still the end goal is to have people effectively apply the math in various domains of application. Knowledge that is relevant to effective application includes mathematical concepts (e.g., data sample, distribution, mean, variance, correlation, independent variables) and mathematical techniques (e.g., Student’s *t*-test, binomial test, regression analysis, analysis of variance, chi-square goodness of fit test), but also, knowledge specifically related to application within the domain. Notably, such application-related knowledge is relevant to *effective* application of probability and statistics to the domain, since the math can also be misapplied (Kuzmak, 2015).

In this paper, a mapping (a cognitive analysis) of knowledge used in the effective application of probability and statistics in real world situations, including both mathematical and non-mathematical knowledge, is presented. Cognitive analysis (within cognitive psychology) is related to other similar analyses of thought and behavior, such as task analysis, job analysis, and knowledge engineering as input to computer modeling of human reasoning and artificial intelligence. Categories of knowledge from the analysis are illustrated using examples of applications involving social data.

Other authors have raised the point that the understanding of how to apply statistics involves more than knowledge of math and calculation. For example, Moore (1998, p. 1253) has noted the broad applicability of statistics across many fields, and likened it to the liberal arts. Wild and Pfannkuch (1999) have proposed models describing statistical thinking in empirical enquiry, including an investigative cycle (PPDAC) with steps: Problem, Plan, Data, Analysis, and Conclusions. In contrast, in this paper, the analysis of the understanding of how to apply statistics is a *mapping of knowledge* involved in the effective application of probability and statistics: and that mapping differentiates mathematical and non-mathematical knowledge. The latter distinction parallels the case of knowledge used for the effective solving of engineering problems, which includes both knowledge of mathematics (e.g., calculus and differential equations) and non-

In: J. Engel (Ed.), *Promoting understanding of statistics about society.*
Proceedings of the Roundtable Conference of the International Association of Statistics Education (IASE),
July 2016, Berlin, Germany. iase-web.org/Conference_Proceedings.php

mathematical knowledge related to application of the math within the domain of engineering. The fundamental role of domain-related knowledge and expertise is reflected in the mapping.

KNOWLEDGE MAPPING

Definitions and Scope

The term *knowledge mapping* is used here meaning a top level analysis intending to provide a complete partitioning of knowledge within a particular scope of interest. Just as a geographic map identifies geographic features (e.g., mountains and rivers) for a terrain of interest, and a political map identifies political divisions (e.g., countries) for a range of territory of interest, similarly, a knowledge map identifies categories of knowledge for a scope of reasoning activity that is of interest. The scope of reasoning of focus here is: the effective application of probability and statistics to real world situations. The phrase "to real world situations" expresses that the focus is on real problems and situations, for which one really seeks and cares about the answer or solution; that is, not just problems that may involve real data, but are unrealistic, oversimplified, or contrived. The scope of knowledge is intended to include the full range of knowledge needed for effective/appropriate application of probability and statistics to real world situations, and, for example, is not limited to a set of fundamental concepts. For simplicity of reference within this paper, this scope of knowledge is herein called *application knowledge*.

Note that a range of professionals are involved in probability and statistics application and education, including educators, academic statisticians, statistical consultants (general), and domain-specific statistics practitioners, as depicted in Figure 1. Considering individuals' backgrounds, some individuals may identify with more than one of the groups arrayed in Figure 1. These professional groups represent a range of perspectives on and degree of familiarity with statistics application and education. With the focus here on effective/appropriate application (within domains of application), domain-related knowledge plays a role, for which domain specialists (represented in three lower sectors of the diagram in Figure 1) can be expected to have the richest experience and knowledge.

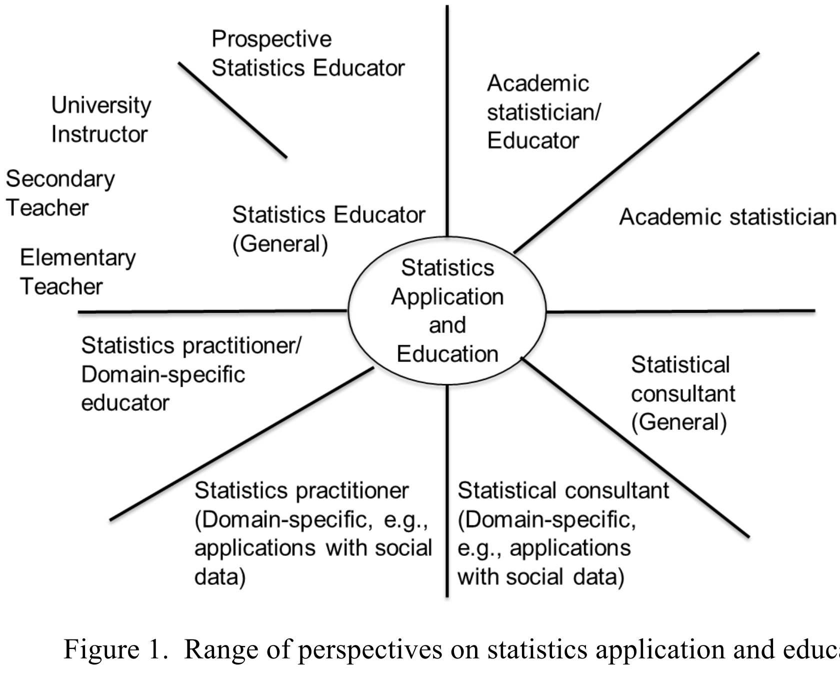

Figure 1. Range of perspectives on statistics application and education

The scope of knowledge that is the focus here spans multiple professional groups and domains of application. At the level of college education, the teaching of the scope of knowledge would span the mathematics/statistics department, together with departments representing domains of application, including social science departments. Domain-specific illustrations provided here are drawn from applications involving social data.

Application Knowledge Mapping

A key feature of a knowledge mapping is a set of categories that provide the framework for the mapping. Just as a political map has categories (such as land vs. water, and country) that provide the organization for mapping the territory of focus, so also a knowledge map is organized by knowledge categories to map the knowledge of focus. The categories provide a partitioning of the scope of knowledge of interest, facilitating the generation of a complete mapping, and providing a basis for determining when the mapping is done. In contrast, an unstructured list of knowledge provides no such facilitation nor basis for determination.

By way of introducing the categories that support the application knowledge mapping of focus here, consider that the effective application of probability and statistics within a domain of application can be viewed as a form of problem solving, wherein the goal is to identify valid statistical evidence relevant to a realistic question that people care about. Through a process of reasoning and problem solving, a solution is produced, with components to the solution. In particular, there is a question (or questions) which is/are addressed, which may be established from the start, or evolve and be established with the benefit of initial analysis. In addition, to address the question, there is a plan to apply statistical techniques, using planned data and design, in a manner that fits real world situation details, e.g., assumptions associated with the technique hold within the application situation. The data supporting the analysis are collected or acquired, ensuring data quality, so that the analysis based on the data will be well-founded. And the analysis is conducted and the results of the analysis interpreted, drawing warranted conclusions. These solution components provide a top level partitioning of knowledge, and are depicted the horizontal blocks of the application knowledge mapping framework in Figure 2.

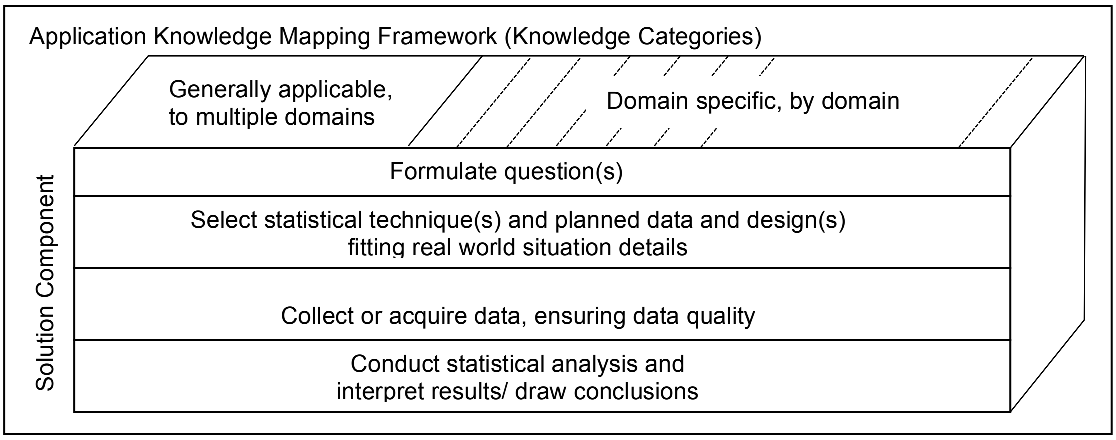

Figure 2. Application knowledge mapping framework, based on knowledge categories, for the effective application of probability and statistics to real world situations

While the solution components in the knowledge mapping framework may appear to reflect a sequential ordering, recall that the intent here is not to describe a cycle of activity, but rather to provide a partitioning of knowledge used in effective application of probability and statistics. In fact, the sequence of execution of components can be expected to vary across different application/ reasoning contexts. For example, exploratory data analysis (Tukey, 1977) may precede formulation of a question that is then the focus of confirmatory analysis. In addition, the extent of activity related to each component will vary in execution, depending on whether the inquiry is original research or has been standardized, as when standard updated reports are produced periodically based on data available from a statistical institute; or depending on whether new data collection is constrained, as occurs for opportunistic mining of big data. Also, intermediate empirical results and discovery during problem solving may influence the course of reasoning, leading to iterations such as refining of the question and planning of additional analysis. Note that knowledge to conduct statistical analysis and interpret results/ draw conclusions are integrated together in the knowledge mapping (in contrast to the PPDAC cycle), since knowledge of statistical techniques includes the knowledge of the nature of conclusions that may be derived through their application, e.g., knowledge of Student's *t*-test includes how to conduct the test, and that results of the test indicate whether or not there is a statistically significant difference for a variable across two samples drawn from two groups/ populations, which leads to inference about any difference between the two groups/ populations.

A main point is that, regardless the sequencing of reasoning, each of the four solution components that comprise the knowledge partitioning in the knowledge map are required for effective application of probability and statistics to real world situations. The variable and iterative nature of the statistical problem solving process is well known to practitioners, and has been noted in the statistics education literature (Konold and Higgins, 2003, p. 194). The knowledge mapping based on required solution components, is compatible with the recognition that the problem solving process is variable in sequence and may include iteration in reaching a solution. Analysis of the cognitive capability that drives the particular sequencing of statistical problem solving (the sequential application of relevant knowledge) is beyond the scope of this paper. General reasoning and problem solving heuristics may apply, such as introduced by Polya (1945/1957).

A second dimension of knowledge categories appears in Figure 2, namely, whether knowledge is generally applicable (to multiple domains of application), or is domain-specific, for a range of domains. This second dimension provides a concurrent basis for partitioning knowledge within the scope of interest. Together, the two dimensions of categorization (solution components and domains of application), provide the framework for the mapping. It then remains, for each block in the framework, to identify and list the knowledge that pertains to that category, to complete the knowledge mapping. By providing a partitioning of knowledge intended to be complete, the categories within the knowledge map provide a guide for evaluating the completeness of instruction in the effective application of probability and statistics.

Just as a geographic map may evolve with exploration and growth in familiarity with the terrain, similarly a knowledge map for a domain of reasoning may evolve in its categories and breadth over time.

Mapping the Knowledge by Category

In the next subsections, the four solution component knowledge categories are described and instances of knowledge in each category are identified, with primary focus on the first vertical slice in Figure 2, namely, knowledge that is generally applicable (to multiple domains). The instances of generally applicable knowledge are illustrated with applications involving social data. In addition, in some cases, instances of domain-specific knowledge related to social data are also identified in the descriptive text. Note that each solution component category involves non-mathematical and domain-related reasoning (not "just math").

Within each knowledge category below, many of the items identified may be familiar to the reader – but recall that the intention here is to provide a systematic mapping or inventory of knowledge, much of which content can be expected to be familiar to a trained practitioner in the application of probability and statistics. However, knowledge that is familiar to the practitioner may not currently be being explicitly taught, instead, for example, be being learned through apprenticeship and personal guidance through the working of many examples. Also, within the mapping, knowledge that is domain-specific can be expected to be less familiar or indeed unfamiliar to the generalist; and may not be being explicitly taught within the instruction for the domain. The descriptions and number of instances of knowledge are limited, given the limited scope of this paper, and may be expanded in future work.

Formulate Question(s)

Questions about what is true in the real world are the focus of statistical investigation. Included in focal questions are references to groups or populations and their attributes; e.g., "What is the city's murder rate, and did it change from the previous year?," "If the U.S. election were held today, which Presidential candidate would win?," or "Which countries have the highest literacy rate?" Knowledge relevant to formulating question(s) for investigation includes:

- *Knowing that definition of terms/ measures is critical.* If sufficient definition detail is not provided, terms/ measures are open to different interpretations and lead to different quantified

results. Term definitions should be sufficiently detailed to ensure reliability of the measure (its reproducibility under like conditions), and should be aligned with how the results are planned to be used. For example, how is "murder" defined? Is it equivalent to murder convictions in court? Are unsolved suspected murders included? Are deaths due to violent civil war included, which occur outside the oversight of an established legal system?
- *Weighing the availability of data, and the cost to collect or acquire data.* Time and cost constraints on an investigation are common, so the availability and cost of data relevant to the question is proper to consider. Which particular data is available and the quantity of data available within the time and cost constraints, may lead one to adjust the focal question(s). For example, one may have data available on the U.S. literacy rate for English from a statistical institute, and be missing data on whether U.S. non-English speakers who are not literate in English are nonetheless literate in their native language.
- *Being mindful of differences between the question(s) of original/true concern, and similar question(s) adopted as a focus of investigation due to data availability and cost factors, so that awareness of the differences are carried through to the interpretation of results.*

Select Statistical Technique(s) and Planned Data and Design, Fitting Real World Situation Details

Based on the focal question(s), statistical technique(s), with appropriate design and planned data, are selected, that address the focal question(s), fitting the real world situation details. Related knowledge includes:
- *Ensuring that assumptions associated with a selected statistical technique fit the real world situation.* For example, use statistical tests that assume normally distributed data (e.g., Student's *t*-test), only when the assumption of normally distributed data is valid for the particular real world situation. Use non-parametric statistical tests as appropriate, e.g., when data distributions are unknown.
- *Using analysis and experimental paradigms that have been established within the application domain, as appropriate.* Statistical techniques are numerous, and investigation opportunities are vast. Within application domains, typically analysis and experimental paradigms have been established that support productive investigation and make appropriate use of statistical techniques. For example, there may be accepted practice within an application domain (justified by the weight of past investigations) regarding the variables for which one should control, in sampling to determine group/ population differences. For example, in public opinion polling, procedures for obtaining representative national samples to support political polling, are well established.

Collect or Acquire Data, Ensuring Data Quality

The value of an investigation is critically dependent on the quality of the data on which it is based. Steps should be taken to ensure data quality; that is, data quality should not be assumed. Related knowledge includes:
- *Evaluating data quality for any prospective data source.* Do not proceed to incorporate data into an investigation unless there is evidence of adequate data quality. For example, evaluate whether there is any divergence between data/ measure definition for an external source, e.g., a statistical institute, compared to the measure/ data definition for the focal question of the investigation.
- *Ensuring any collected data matches the measure/ data definition details associated with the focal question.* If person(s) are involved in data collection for the current investigation, ensure that each is well trained in the measure/ data definitions and data collection procedures. If multiple persons are involved, evaluate the consistency of data collection performance across persons (i.e., interrater reliability).
- *Reviewing the collected/ acquired data using visual/ graphical representation, identifying any outlier data, and evaluating whether the outliers are errors or valid data.* Include or drop outlier data, as there is justification.

The importance and elusiveness of knowledge in this overall category has been previously acknowledged, summarized as "The importance of data production" in the GAISE College Report

(2005, p. 8), and attributed to G. Cobb (1992). However, the GAISE College Report (2005) does not elaborate the knowledge detail within this category such as the three items above. In the Executive Summary of the Report (p. 5), the recommendation appears to "Use real data," without qualification regarding data quality; and the more detailed discussion of this recommendation (pp. 16-17) emphasizes the value of real data in engendering students' interest and motivation, without emphasis on data quality. More recently, the GAISE College Report (Draft February 2016), has relegated mention of "The importance of data production" to an Appendix (p. 28), and the earlier description of this phrase, present in the 2005 GAISE College Report, including the connection to data quality, has been dropped from the draft report.

Conduct Statistical Analysis and Interpret Results/ Draw Conclusions

The value of an investigation is critically dependent on the interpretation of results being warranted, soundly following from the statistical analysis and related reasoning. Related knowledge includes:

- *Addressing any differences between the question(s) of original/true concern, and similar question(s) adopted as a focus of investigation due to data availability and cost factors.* The differences may be addressed by estimating the answer to the original question based on the answer to the adopted question, and addressing confidence in the estimate if possible; or by simply noting and discussing the error inherent in using the answer to the adopted question as an estimate of the answer to the original question. Avoid wishful thinking and glossing over the differences, and treating the analysis as if it has answered the original question, when it has not.
- *Addressing statistical significance vs. practical significance.* For large samples sizes (e.g., greater than 200), even small observed differences between sample means (averages) for samples from two groups/populations turn out to be statistically significant. However, such small observed differences typically make no practical difference, not warranting any change in how the two sampled groups/populations are viewed or treated. For example, an increase in a country's literacy rate from one year to the next may be statistically significant, but yet be a small percentage increase and practically insignificant, not newsworthy.
- *Avoiding overgeneralization of results.* Although the focal question and data analyzed in an investigation may be limited in scope (e.g., a study of Presidential candidate preferences of students at a particular college in the U.S.), there may be a tendency to overgeneralize investigation findings (e.g., unjustifiably to conclude that all U.S. college students have the same pattern of candidate preferences as was observed for the one college, or that all U.S. citizens have that same pattern of candidate preferences). Such overgeneralizations, which discount the variation that exists across the population, should be avoided.

IMPLICATIONS FOR EDUCATIONAL PRACTICE

College and High School Level

The knowledge mapping and examples above illustrate that both mathematical and non-mathematical/ domain-related knowledge are required for effective application of probability and statistics within a domain of inquiry. Thus, if only "the math" is taught, instruction for effective application of the math is incomplete. The knowledge mapping provides a basis for evaluating the completeness of current curriculum and instruction for applied probability and statistics, and a guide for designing more complete curriculum and instruction. Providing more complete instruction with respect to addressing the full range of required knowledge, may address widespread and persistent problems that have been observed with the effectiveness of current teaching and learning of probability and statistics (Garfield & Ahlgren, 1988; Garfield & Ben-Zvi, 2007; Tishkovskaya & Lancaster, 2012).

Given that one accepts the goal of improving the addressing of the full range of required knowledge for effective application of probability and statistics, the issue arises of how to accomplish that goal at the university level. Mathematics departments typically see their scope as teaching "the math," and see the teaching of non-mathematical/ domain-related knowledge relevant to applying statistics in various domains as out of their scope. University departments for fields that use statistical inquiry, such as social science departments, typically have their own courses on applied statistics, generally modeled after the math courses in that the focus in on teaching statistical concepts and techniques, with the benefit of emphasis on techniques useful in the domain, and application examples drawn from the domain; but still may include little explicit instruction on relevant non-mathematical/ domain-related knowledge such as how to ensure data quality within the domain.

To teach effective application of probability and statistics, calls for a coordinated solution across university departments, with that solution collectively addressing the full range of relevant knowledge. Such a coordinated solution for teaching the effective application of probability and statistics, using the application knowledge mapping, is summarized in Figure 3.

Allocation of Teaching of Knowledge for the Effective Application of Probability and Statistics

Mathematics/ Statistics Dept.	Social Science Dept.
• Mathematical foundations: Probability, Statistics • Overview of Application Knowledge Mapping, illustrated with selected real examples from a range of domains • Application Knowledge that is generally applicable, across multiple domains.	• Statistical concepts and techniques used in the Social Science domain • Overview of Application Knowledge Mapping, illustrated with real examples from the Social Science domain • Application Knowledge that is generally applicable, across domains; and domain-specific application knowledge, illustrated with real examples from the Social Science domain

Figure 3. Allocation of knowledge for the effective application of probability and statistics to be taught among university departments

First, within the mathematics or statistics department, while the course focus may be on the mathematics of probability and statistics, since the power and value of statistics lies in its appropriate application, the subject of effective application of statistics within domains and required non-mathematical/ domain-related knowledge, should also be raised. An overview of that required application knowledge should be presented (described in the application knowledge mapping above), as well as presenting the instances of application knowledge generally applicable across domains (e.g., concerning the collection or acquisition of data, ensuring data quality). Effective application of probability and statistics should be illustrated with examples drawn from a variety of domains, having ensured that the examples meet normative standards within the domain based on consensus of experts (Kuzmak, 2015). Inclusion of such examples not only serves to illustrate the value of probability and statistics through statistical inquiry, but also serves to demonstrate to the student that there is a role of non-mathematical and domain-related knowledge in ensuring effective application of probability and statistics. This addresses the issue that mathematics/ statistics instruction, with underemphasis on application issues, may effectively teach students to misapply statistics (Kuzmak, 2015).

University departments for fields that use statistical inquiry should consider explicit inclusion in instruction of more of the non-mathematical/ domain-related knowledge required for effective application of probability and statistics within their field, as summarized in Figure 3. While some of this knowledge may currently be included in instruction within the department, the knowledge mapping above provides a guide for evaluating the completeness of instruction and identifying opportunities to be more complete. More explicit instruction in the required non-mathematical/ domain-related knowledge can be expected to advance the practice of effective application of probability and statistics.

An additional general benefit of the application knowledge mapping is that, by providing a focal point for enhanced collaboration and coordination among university departments, opportunities may be identified to reduce unnecessary redundancy in course content across course offerings and achieve increased educational efficiency.

All Grade Levels

The application knowledge mapping provides a resource to review and evaluate curriculum in statistics and data science at all grade levels to ensure misapplication of probability and statistics is not taught. To illustrate this point, consider the following class exercise from the perspective of modeling effective/appropriate application of probability and statistics to real world situations involving social data.

> *In a class of 26 college students, students are asked to pair up, and to interview each other to ask each other their planned major and record their response. The data on each student's planned major are pooled together. A bar chart showing the frequency of planned majors for the class is generated.*

Does the class exercise above model effective/appropriate application of probability and statistics in real world situations, regarding formulating the question? Ensuring data quality? Planning statistical analysis?

The application knowledge mapping, including knowledge generally applicable (for multiple domains) pertaining to required solution components, provides a basis for such evaluation. Considering the knowledge in the application knowledge mapping reviewed above, one may ask, is question formulation absent or unclear? Is there a lack of attention to data quality? This is especially an issue for social data, which may be derived from interviewing or self-reporting, and may involve multiple data collection personnel. Is there a lack of planning and coherence among question formulation, data collection and acquisition, and data analysis? Affirmative answers to these questions are indicative that there is a risk that misapplication of probability and statistics is being taught, and this is so for the class exercise above.

To address the risk of teaching misapplication of probability and statistics, one needs to provide clarification regarding the valid lessons learned for exercises illustrating concepts in probability and statistics, and include caveats and cautionary notes, when necessary activities for real world application of probability and statistics have been skipped or oversimplified.

CONCLUSION

Effective application of probability and statistics to real world situations is addressed by providing a knowledge map for related knowledge, intended to provide a complete partitioning of knowledge, including mathematical and non-mathematical/ domain-related knowledge. It is noted that teaching "just the math," with underemphasis on application issues, may effectively teach students to misapply statistics (Kuzmak, 2015). The knowledge map is proposed as a guide to evaluate the completeness of current curriculum and instruction in applied probability and statistics, to design more complete and effective curriculum and instruction, to coordinate curriculum design across university departments, and to evaluate current curriculum at all grade levels to ensure misapplication of probability and statistics is not taught.

REFERENCES

Aliaga, M., Cobb, G., Cuff, C., Garfield, J., Gould, R., Lock, R., Moore, T., Rossman, A., Stephenson, B., Utts, J., Velleman, P., & Witmer, J. (2005). *Guidelines for Assessment and Instruction in Statistics Education (GAISE) College Report.* American Statistical Association. Retrieved from www.amstat.org/education/gaise/

Carver, R., Everson, M., Gabrosek, J., Rowell, G. H., Horton, N., Lock, R., Mocko, M., Rossman, A., Velleman, P., Witmer, J., & Wood, B. (2016). *Guidelines for Assessment and Instruction in Statistics Education (GAISE) College Report (Draft February 2016).* American Statistical Association. Retrieved from www.amstat.org/education/gaise/

Garfield, J. & Ahlgren, A. (1988). Difficulties in Learning Basic Concepts in Probability and Statistics: Implications for Research. *Journal fo r Research in Mathematics Education, 19*, 44-63.

Garfield, J. & Ben-Zvi, D. (2007). How Students Learn Statistics Revisited: A Current Review of Research on Teaching and Learning Statistics. *International Statistical Review, 75*, 372-396.

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., & Kruger, L. (1989). *The Empire of Chance.* Cambridge, UK: Cambridge University Press.

Konold, C. & Higgins, T. (2003). Reasoning about Data. In J. Kilpatrick, W. Martin, & D. Schifter (Eds.), *A Research Companion to Principles and Standards for School Mathematics*, 193-215. Reston, VA: National Council of Teachers of Mathematics (NCTM).

Kuzmak, S. (2015). A Cognitive Framework for Normative Reasoning under Uncertainty, and Reasoning about Risk, and Implications for Educational Practice. *The Mathematics Enthusiast, 12*(1,2,& 3), 140-156.

Moore, D. (1998). Statistics among the Liberal Arts. *Journal of the American Statistical Association, 93*, 1253-1259.

Polya, G. (1957). *How to Solve It*, (Second edition). Princeton, N.J.: Princeton University Press. (Original work published 1945)

Tishkovskaya, S. & Lancaster, G. (2012). Statistical Education in the 21st Century: A Review of Challenges, Teaching Innovations and Strategies for Reform. *Journal of Statistics Education, 20*(2). Retrieved from www.amstat.org/publications/jse/

Tukey, J. (1977). *Exploratory Data Analysis* (First edition). Reading, MA: Addison-Wesley.

Wild, C. & Pfannkuch, M. (1999). Statistical Thinking in Empirical Inquiry. *International Statistical Review*, 67, 223-265.