ASSESSMENT OF INFORMAL AND FORMAL INFERENTIAL REASONING: A CRITICAL RESEARCH REVIEW

MARIA GUADALUPE TOBÍAS-LARA
Tecnológico de Monterrey
mgtl@itesm.mx

ANA LUISA GÓMEZ-BLANCARTE
Instituto Politécnico Nacional, CICATA-Legaria
algomezb@ipn.mx

ABSTRACT

As a contribution to the discussion on the assessment of informal inferential reasoning (IIR) and the transition from this to formal inferential reasoning (FIR), we present a review of research on how these two types of inferential reasoning have been conceptualized and assessed. Based on our review, we discuss the need to redefine the conceptions of IIR and FIR in order to create an integrated description of inferential reasoning that includes not only ideas of IIR and FIR, but also the whole activity of argumentation, which involves the production of both statistical and contextual reasons. Current descriptions of IIR and FIR list the facts that might be brought from data analysis to the process of inferential reasoning. The approach we propose considers how the facts, both statistical and contextual, can be used as arguments, leading to assessments of students' inferential reasoning focusing on articulating the statistical and contextual reasons students present to support an inference.

Keywords*: Statistics education research; Statistical inference; Statistical and contextual arguments; Assessments of students' reasoning*

1. INTRODUCTION

Today, statistical inference is applied in diverse fields where evaluation of information must be grounded in data-based evidence (Johnson, 2011). Unfortunately, research has shown that statistical inference is frequently taught as an isolated topic, instead of being (1) clearly integrated with other statistical concepts and exploratory data analysis (Rossman & Chance, 1999), or (2) connected to research frameworks or experimental design (Batanero, 2000). There is a second issue discussed by Batanero around the questions of "when, how, and how deeply we should teach statistical inference" (Batanero, 2000, p. 15). It is often simply assumed that students have assimilated the key statistical concepts needed to comprehend statistical inference when, in reality, this level of understanding is not always achieved (Vallecillos & Batanero, 1997).

This deeper appreciation of the difficulties entailed in teaching and learning statistical inference led to educational reforms that invited educators to focus more on statistical concepts, relationships among them, and the development of statistical thinking (Cobb & Moore, 1997; Moore, 1997). Derived from these reforms, special interest in the study of statistical inferential reasoning arose within the statistics educational research community. Since then, research on inferential reasoning has grown substantially, covering several key aspects that have helped improve our understanding of the teaching and learning of statistical inference. For example, this type of reasoning has been a topic of discussion at international events like the Fifth International Forum on Statistical Reasoning, Thinking and Literacy (SRTL-5) and the 9th International Conference on Teaching Statistics (ICOTS-9), and has been a theme for special issues in research journals, including the *Statistics Education Research Journal* (volume 7, number 2), *Mathematical Thinking and Learning* (volume 13, numbers 1 and 2), and *Educational Studies in Mathematics* (volume 88, number 3). Finally, it is the focus of a chapter of the book *Learning to Reason about Statistical Inference* by Garfield and Ben-Zvi (2008).

Statistics Education Research Journal, 18(1), 8–25, http://www.stat.auckland.ac.nz/serj

Park (2013) defined inferential reasoning in the field of statistics as "the reasoning that people use when drawing conclusions from data" (p. 245). Currently, research by the statistics education community classifies inferential reasoning into two types: *informal* and *formal* inferential reasoning. Assuming that formal ideas of statistical inference should be constructed progressively while gradually increasing the level of formalization (Batanero, 2011; Garfield & Ben-Zvi, 2008; Zieffler, Garfield, delMas, & Reading, 2008), researchers have focused more on the study of informal inferential reasoning (IIR). In general, the main idea shared by the research community in statistics education is that promoting IIR should be emphasized with students, especially children and teenagers, because of the support they require to better understand the foundations of statistical inference. Informal inference has been used as a teaching strategy to address certain difficulties in learning statistical inference. As Wild, Pfannkuch, Regan, and Parsonage (2017) observed, "many of the problems with students learning statistics stem from too many concepts having to be learned and operationalised almost simultaneously, pointing to the need for complexity-reduction strategies" (p. 105). Some researchers suggest that IIR may provide a pathway to achieve more formal inferential reasoning (e.g., Garfield, Le, Zieffler, & Ben-Zvi, 2015; Weinberg, Wiesner, & Pfaff, 2010). For instance, in their work, Garfield et al. (2015) express an interest in "the development and use of IIR to build formal statistical inferential reasoning" (p. 334), while Teng (2016) observed that "the concept of informal inferential reasoning (IIR) was conceived as a cognitive tool to pave the way for formal inferential reasoning" (p. 87).

One of the current challenges in research on inferential reasoning involves how to study its cognitive development (Reading, 2007), so that we can determine how IIR evolves into formal inferential reasoning (FIR). For example, some researchers have proposed analyzing more deeply the best sequences of ideas and activities and their role in the transition from informal to formal methods of statistical inference (Zieffler et al., 2008). Jacob and Doerr (2013), meanwhile, recognized the need to examine how IIR is developed and how students can demonstrate it in a variety of contexts. Garfield and Zieffler (2011) contributed by suggesting research questions that include "how does good informal inferential reasoning foster students' ability to use and understand formal methods of statistical inference?" (p. 280). For these authors, "a key question is whether good informal inferential reasoning can stand alone, or if it must be a stepping-stone to formal methods of inference" (p. 280). Referring to studies of IIR at the tertiary level, "research has been useful in describing students' reasoning at this level, but more research is needed to provide evidence of whether and how informal statistical inference improves the transition to formal statistical inference, as has been hypothesised" (Makar & Rubin, 2014, p. 4).

With the aim of contributing to the discussion on the cognitive development of IIR and the transition to FIR, this article presents a critical review of research that has evaluated both approaches to reasoning. The study is organized around three key issues: (1) the characteristics of IIR and FIR, (2) assessments of IIR and FIR, and (3) an evaluation of the relationship between IIR and FIR. The first section provides an overview of the characteristics, or components, that distinguish formal from informal inferential reasoning; the second explains how IIR and FIR have been evaluated, indicating the aspects of these approaches to reasoning upon which researchers have focused attention, while the third describes proposals for assessing a possible relationship between IIR and FIR. In the final section, we discuss, first, how IIR and FIR are currently conceptualized and, second, how the context and statistical aspects of reasoning have been considered in relation to assessing students' inferential reasoning. In so doing, we highlight the role that both *statistical* and *contextual* reasons play in inferential reasoning, and suggest ways to assess students' inferential reasoning.

2. METHODS

The literature review was conducted using the documentary research technique; that is, we analyzed research papers pertaining to the following keywords, which could appear anywhere in the paper: statistical inference, informal inferential reasoning, formal inferential reasoning, informal statistical inference, and formal statistical inference. The sources consulted included statistics education journals (e.g., *Journal of Statistics Education, Statistics Education Research Journal, International Statistical Review, Journal of the Royal Statistical Society*) and mathematics education journals (e.g., *Mathematical Thinking and Learning, Educational Studies in Mathematics, RELIME, The College Mathematics Journal, ZDM International Journal on Mathematics Education*). In addition, the

proceedings of congresses on statistics education (e.g., International Conference on Teaching Statistics) or mathematics education (e.g., International Congress on Mathematical Education), books, book chapters, and doctoral dissertations were reviewed. All the studies compiled were written between 1990 and 2016 in English or Spanish. A total of 82 sources were identified: 64 studies of IIR, 10 reports on FIR, and 8 papers that dealt with both types of reasoning. To select the most important studies, we established inclusion criteria according to the focus of this article (see Table 1).

Table 1. Inclusion criteria and the focus of the review

Inclusion criteria	Review focused on
	Issue 1: Characteristics of IIR and FIR
Original works that provide meanings and characteristics of components that make it possible to distinguish IIR from FIR. Hence, we discarded articles that contained such information, but only cited original sources.	1. The meaning of both types of reasoning: IIR and FIR. 2. The characteristics or components that define them.
	Issue 2: Assessment of statistical inferential reasoning
Sources with information on *how* IIR and FIR have been evaluated and *what* has been assessed were included.	1. How inferential reasoning (informal and formal) was evaluated (e.g. tests, theoretical models). 2. What aspects of inferential reasoning were evaluated (e.g., concepts of inference, data exploration, descriptive statistics). 3. What aspects of inferential reasoning were the focus of the evaluation (e.g., definitions, applications, conceptual ideas, understandings, interpretations, or procedures).
	Issue 3: Assessing the relationship between IIR and FIR
Sources that assess a relationship or association between the two types of inferential reasoning (formal and informal).	1. How the relationship between IIR and FIR was assessed (e.g., tests, theoretical models). 2. What issues of both types of inferential reasoning (formal and informal) were assessed. 3. What aspects of the association or relationship between the two types of inferential reasoning were shown.

Of the 82 sources identified, 33 were analyzed critically in relation to the three issues examined in this article, after fulfilling the established inclusion criteria (Table 2).

Table 2. Summary of the sources that fulfilled the inclusion criteria

Type of paper	Studies on IIR	Studies on FIR	Studies on IIR and FIR
Journal	15	4	1
Proceedings	6	2	0
Book chapter	0	0	1
Doctoral dissertation	1	0	3

3. CHARACTERISTICS OF IIR AND FIR

This section presents an overview of how IIR and FIR have been distinguished.

3.1. FORMAL INFERENTIAL REASONING

Makar and Rubin (2009) explained that "by formal statistical inference, we refer to inference statements used to make point or interval estimates of population parameters, or formally test hypotheses (generalizations), using a method that is accepted by the statistics and research community" (p. 85). They further maintained that statistical inference includes both an outcome and the reasoning

applied to create probabilistic generalizations from data. Park (2013), meanwhile, referred to formal statistical inference as a content domain that includes concepts and ideas related to formal, standard statistical testing (e.g., confidence intervals, hypothesis testing, *p*-value). In addition to confidence intervals, other methods of statistical inference considered "formal" are tests of significance and regression models (Pfannkuch, 2006a).

The few studies that we found on FIR indicated that this reasoning is associated with the use of formal methods and procedures, such as *p*-value, hypothesis testing, confidence intervals (Weinberg et al., 2010; Zieffler et al., 2008), formulas, calculations, and statistical tables that are used to draw formal statistical inferences (Huey, 2011). Regarding FIR, Huey observed that students need to acquire not only knowledge of formal statistical methods, but also the understanding of when to use them. Huey stated that the goal of teaching FIR is to ensure that students "are able to inferentially [sic] reason with formal methods in a meaningful way and answer questions about why the processes are effective, how they work, and what the results do and do not imply" (p. 11).

3.2. INFORMAL INFERENTIAL REASONING

The informal approach to statistical inference comprises a much broader set of competencies because it is not limited exclusively to formal procedures. Informal inferences are uncertain generalizations that require a reasoned process to extend the conclusion beyond data description, and are open to considering statistical inference outside formal methods (Makar, Bakker, & Ben-Zvi, 2011; Makar & Rubin, 2009; Park, 2013). Thus, IIR is related to the following: students' intuitions when faced with inferential tasks, the use of descriptive analyses of sample data, exploratory data analysis, and simulations, all of which are considered preliminary versions of formal inference in which the conceptual ideas of statistical inference are not yet fully developed. The purpose of this type of inference is to introduce inferential ideas before they are seen formally (Garfield & Ben-Zvi, 2008; Garfield, delMas, & Zieffler, 2012; Pfannkuch, 2011; Zieffler et al., 2008). For example, exploratory data analysis can be based on interpretations of graphic representations, analyses of statistical measures, statistical models generated with random sample data (Ben-Zvi, Gil, & Apel, 2007), which make it possible to visualize the characteristics of data distributions (Pfannkuch, 2006a, 2006b), models that simulate randomness (Rossman, 2008), and estimates of uncertainty (Makar, 2013). Reading (2007) pointed out that developing IIR requires understanding such basic ideas as variation, distribution, median, dispersion, and interpretation of graphs, and then presented four statistical actions identified by research as necessary building blocks for IIR: seeing data as an aggregate, focusing on proportions instead of absolutes, appreciating the variability in samples, and understanding randomization as a reasoning process.

Researchers have described IIR using different language, as the following citations show:

- "the way in which students use their informal statistical knowledge to make arguments to support inferences about unknown populations based on observed samples" (Zieffler et al., 2008, p. 44)
- "the use of informal statistical knowledge to estimate inferences about unknown populations wherein generalizations beyond data are made, grade-level appropriate analysis of data is used as evidence, and probabilistic language in expressing uncertainty about generalizations is present" (Goss, 2014, p. 100)
- "including reasoning about sample and sampling as well as reasoning about informal inference" (Gil & Ben-Zvi, 2014 p. 1)
- "how students use their knowledge to make and sustain statistical inferences on an unknown population based on observed samples [but] without utilizing the formal methods or techniques of inferential statistics, such as sampling distribution, standard deviation, standard scoring, etc." (García & Sánchez, 2014, p. 348)
- "the reasoning processes leading to informal statistical inference" (Ben-Zvi, Aridor, Makar, & Bakker, 2012, p. 916)
- "the process of drawing conclusions, making interpretations, making judgment or decisions about populations based on representations of data, or statistical summaries of data that came from samples" (Nor & Idris, 2010, p. 4805)

- "accomplished by comparing two or more aspects of data when generating an inference without the assistance of a formal algorithm" (Huey, 2011, p. 3).

Zieffler et al. (2008) and Makar and Rubin (2009) identified key components of IIR (see Table 3) and proposed that it can be studied on that basis.

Table 3. Key components of IIR

Zieffler et al. (2008)	Makar and Rubin (2009)
1) Judgments, predictions or statements made about the population based on the analysis of sample data, without using formal methods.	1) Generalizations that can be predictions, parameter estimations, and conclusions, extend beyond the data of the samples to the population, and generally made by identifying data patterns.
2) The formal and informal prior knowledge used in the reasoning process to make an informal inference. This prior knowledge could include context, fundamental concepts, statistical language, and informal knowledge of inference.	2) The evidence used to justify those generalizations must be based on statistical data, and must be accepted within the context in which they are used.
3) Articulating evidence arguments when drawing a judgment, prediction or declaration of the population based on samples. It refers to the statistical evidence that is presented in explanations that support the conclusion.	3) Employment of probabilistic language (uncertainty) when articulating or writing an inferential generalization, prediction, or conclusion. This language refers to uncertainty levels about the conclusions drawn.

Clearly, there is no single, unique definition of IIR, although all these articulations recognize that it does not utilize formal methods of statistical inference (e.g., confidence interval, hypothesis tests), but that its use makes it possible to introduce and construct more formal ideas of statistical inference. Like Garfield et al. (2015), we consider the key IIR components (Table 3) as not exclusive to IIR, but also operate in FIR; that is, both types share similar characteristics because they (1) are generalizations, judgments, or statements that go beyond sample data, (2) require statistical evidence to support a conclusion or generalization, (3) employ probabilistic language, and (4) demand knowledge of the context in which the inference takes place.

3.3. DIFFERENCES AND SIMILARITIES BETWEEN IIR AND FIR

Makar et al. (2011) affirmed that in IIR "we are not expecting students to rely on formal statistical measures and procedures to formulate their inference" (p. 153). In this sense, one aspect that seems to differentiate between the two types of inferential reasoning consists of the statistical methods used to generate evidence for the inference (see Table 4). We, however, consider the process of inferential reasoning to go beyond formal or informal statistical measures and procedures, because "what is informal could depend on the nature of the inferential tasks being studied, on the complexity of the statistical or probabilistic concepts involved, on the educational stage, and on other factors" (Pratt & Ainley, 2008, p. 3). Moreover, according to Bakker, Kent, Derry, Noss, & Hoyles (2008) and Bakker

Table 4. Similarities and differences between IIR and FIR

Similarities	Differences
Both types of inferential reasoning 1) are generalizations, judgments, or statements that go beyond sample data, 2) require statistical evidence to support a conclusion or generalization, 3) employ probabilistic language, 4) demand knowledge of the context in which the inference takes place.	1) Data analysis in FIR relies on formal methods of statistical inference (e.g., confidence intervals, hypothesis testing, *p*-value, tests of significance, regression models, sampling distribution, standard deviation, standard scoring). 2) Data analysis in IIR are open to considering the use of descriptive analysis of sample data, exploratory data analysis, simulations, and uncertainty estimates.

and Derry (2011), the way in which a person reasons depends, as well, on a set of techniques and norms that are usually accepted as formal in certain fields of science in which inference takes place, although in statistics those same techniques or norms could be considered informal, or *vice versa*. Other authors (e.g., Garfield et al., 2015), meanwhile, pointed out that an expert statistician can reason both formally and informally, depending on the nature of the inferential problem and the type of decision required.

According to Ben-Zvi (2006), inferential reasoning is comparable to argumentation, as there is a need to construct persuasive arguments based on data analysis, which can be formal or informal. In this way, we think that informal and formal inferential reasoning are two different methods of argumentation to support a given inference. From a point of view of Bakker et al. (2008), we consider these methods to be *statistical reasons*, which are only an organic part of the *space of reasons* in which statistical inference takes place, especially in workplace situations. A statistical inference should take place in a space of reasons that "encompasses what can be analytically distinguished as *contextual* and *statistical* reasons" (Bakker et al., 2008, p. 139). Contextual reasons refer to norms related to the context of the problem in which the statistical inference is applied—or real-world constraints—whereas statistical reasons are derived from the results of data analysis (Bakker et al., 2008). Therefore, both informal and formal inferential reasoning take place in that space of reasons. "Contextual or statistical reasons are prioritised depending on whatever is required to reach a goal" (Bakker et al., 2008, pp. 139–140).

4. ASSESSMENT OF STATISTICAL INFERENTIAL REASONING

Researchers have determined the level of student reasoning in both FIR and IIR mainly by identifying the characteristics of the reasoning. The Structure of the Observer Learning Outcome (SOLO) taxonomy and some quantitative tests have been utilized to carry out such evaluations. The next section describes how FIR and IIR have been evaluated, the aspects of these forms of reasoning that have been assessed, and the specific focus of the assessments.

4.1. ASSESSMENT OF FORMAL INFERENTIAL REASONING

In the literature reviewed, we found that FIR assessments focused on ascertaining students' thinking about formal concepts such as hypothesis testing, *p*-value, statistical significance, confidence intervals, sampling distribution, and significance level, as well as on interpretations of these concepts and the reasoning process involved in a hypothesis test or a confidence interval (e.g., Huey, 2011; Inzunsa & Jiménez, 2013; Jacob, 2013; Lane-Getaz, 2013; Park, 2012, 2013).

The test designed by Lane-Getaz (2007), called Reasoning about P-values and Statistical Significance (RPASS), was originally designed for assessing statistical literacy of *p*-value and statistical significance. Other formal concepts of statistical inference (hypothesis testing, Type-I and Type-II error rates), and confidence intervals were included in the version 7 of the RPASS test. Using the RPASS-7 test, Lane-Getaz (2013) evaluated 105 college students on the following learning outcomes: (1) defining a *p*-value, (2) using tests of significance, (3) interpreting results through hypothesis testing, (4) drawing conclusions about statistical significance, (5) tying *p*-values back to hypotheses (which entails linking the *p*-value to the conclusions of hypothesis testing), and (6) making inferential connections. For each outcome, the RPASS-7 presented one or two problem scenarios that included a set of associated items of different types: true-false, multiple choice, valid or invalid reasoning, valid or invalid conclusion, valid or invalid interpretation, and/or valid or invalid action.

Lane-Getaz (2013) proposed that different interpretations can be developed when making a statistical inference, with the aim of identifying the conceptual ideas of statistical inference that students could generate. For example, two items evaluated formal and informal ideas of *p*-value. For Lane-Getaz, the informal idea of *p*-value was related to the proportion of the empirical data that is at least as unusual as the observed value of a statistic, if the null hypothesis is true, whereas the formal idea was the probability that the result of the statistic would be at least as extreme as the one observed in a random sample under the assumption that the null hypothesis is true. To evaluate students' understanding of these ideas, Lane-Getaz used problems in which, given a certain *p*-value, students had to determine whether the way in which it was interpreted was correct or incorrect. In other problems, students had to estimate the *p*-value based on the distribution of the statistic; that is, empirical (in the informal idea) or theoretical (in the formal idea) distribution. Lane-Getaz also evaluated whether the results of hypothesis

testing were correctly classified as statistically significant or statistically non-significant, according to the result of the *p*-value. The RPASS-7 test includes problems designed to observe whether students can differentiate between statistical significance and practical significance, as well as others that evaluate whether they detect the effect of a small sample size on the inference.

In summary, RPASS-7 focuses on assessing conceptual ideas (e.g., *p*-value, statistical significance, confidence interval, hypothesis testing, significance level), interpreting results in hypothesis testing, identifying the validity or invalidity of a reasoning, an action or an interpretation of a concept or conclusion, making connections between concepts in hypothesis testing, and recognizing the necessary conditions in study design that must be satisfied to conduct hypothesis testing (Lane-Getaz, 2013).

Similar to Lane-Getaz, Park (2012, 2013) created an instrument to evaluate the inferential reasoning of students enrolled in introductory statistics courses at the university level. The instrument is a test called Assessment of Inferential Reasoning in Statistics (AIRS), and it presents items in a multiple-choice format covering the following aspects of inferential reasoning: informal inference, sampling distributions, and formal inference (Park, 2012). The final version of AIRS was administered to 2,056 university students at different institutions in the USA, 1978 of whom completed it. One of the conceptual ideas assessed in Park's (2012) instrument was related to the *p*-value (informal and formal ideas). In this regard, Park observed the conceptions of these ideas that students exhibited when faced with contextualized inferential problems. We perceive that Park shares the same definition of informal *p*-value as Lane-Getaz, because one of the AIRS test questions notes a *p*-value as the proportion of the simulated data that is equal to, or more extreme than, a certain value of the statistic when considering the null hypothesis to be true. Lane-Getaz (2013) and Park (2012) also agreed on the importance of assessing interpretations of the conclusion in the context of the inferential problem. Whereas Lane-Getaz focused on identifying whether students appreciate the difference between statistical and practical significance given the context of the inferential situation, Park sought to assess whether practical significance is integrated as part of various aspects (e.g., sample size, data quality, effect size, method selection, random assignment) that should be included when interpreting the result of hypothesis testing.

To evaluate reasoning related to formal inference, AIRS focuses on assessing the foundations of the formal inference (e.g., samples and sampling, the law of large numbers, population distribution, frequency distribution, the central limit theorem) and aspects related to formal inference, such as the definition, role, and logic of hypothesis testing, conceptual definitions, and conceptual interpretations (e.g., *p*-value and statistical significance, confidence interval, hypothesis testing), connections among concepts like the effect of sample size on the *p*-value and statistical significance, interpretation of the result of hypothesis testing, the choice of the appropriate statistics in hypothesis testing, and recognition of the necessary conditions of the study design, like understanding that an experimental protocol with random assignment supports causal inference.

Some of the formal inferential concepts mentioned above were evaluated by Jacob (2013) who analyzed 136 students' FIR (grades 11 and 12) at two different schools. All students took part in three informal statistical inference classroom activities during a school year and then seven pairs of students took part in four task-based interviews, one of them related to formal inferential tasks (confidence intervals). At the end of the course, all students answered a post-test assessment containing 22 questions: 10 on formal statistical inference. The 10 questions assessed students' FIR related to the interpretation of a confidence interval (three questions), the relationship between sample size and the size of the confidence interval (one question), the meaning of a 90% confidence interval (one question), the interpretation of the *p*-value (four questions), and drawing conclusions based on the results of hypothesis testing (one question).

In contrast to the evaluation methods proposed by Lane-Getaz (2013), Jacob (2013), and Park (2012), Inzunsa and Jiménez (2013) applied the SOLO model to characterize the justifications that 11 university-level mathematics students provided upon answering a questionnaire with 10 questions designed to assess reasoning during hypothesis testing. The format of the questions varied (open, true-false, multiple choice), but in all cases students were required to justify the response they chose. Specifically, Inzunsa and Jiménez evaluated the reasoning of hypothesis testing in relation to the following aspects: (1) the global logic of the hypothesis testing process, (2) the definition of the significance level, (3) the formulation of hypotheses, (4) the relationship of the *p*-value to sample size and statistical significance, (5) test processes using the Student's *t* distribution, and (6) testing processes

with paired data samples. These authors concurred with Park (2012, 2013) on the importance of assessing whether students can choose the appropriate hypothesis test procedure from the study context.

When we scrutinized the problems posed in the questionnaire administered by Inzunsa and Jiménez (2013), we noted that, in general, they focused on evaluating the interpretation of the results of hypothesis testing, whether students chose the adequate hypothesis test procedure according to the inferential situation, whether they could adequately apply the procedure, conceptual interpretations (e.g., *p*-value, significance level, hypothesis testing), hypothesis formulation according to the problem, and the inter-connection of concepts, such as the *p*-value result with statistical significance, or the effect of sample size on the *p*-value.

The SOLO taxonomy was also used by Huey (2011) to characterize 33 pre-service teachers' reasoning about formal inferences. In constructing the FIR characterization, Huey identified three components: (1) appropriate selection of the formal method or hypothesis testing, (2) inference methods executed correctly, and (3) inference result interpreted reasonably. Like Park (2012, 2013) and Inzunsa and Jiménez (2013), Huey assessed the appropriate selection of the hypothesis test procedure as one component of the teachers' FIR.

During a course that included descriptive statistics, probability and probability distributions, sampling and estimation, confidence intervals, and hypothesis testing, Huey (2011) assessed the teachers' inferential reasoning using three instruments: a pre-assessment (applied to all 33 teachers on the first day of class), clinical interviews (held with a sample of 12 teachers at mid-course), and a post-assessment (applied to all 33 teachers in the last week). Between pre-assessment and the interview, the teachers received instruction on descriptive statistics, probability, probability distributions, and sampling. After the interview, but prior to post-assessment, they received classes on methods of formal inference. The responses to the three assessment instruments were classified according to the SOLO taxonomy—Pre-structural, Uni-structural, Multi-structural, and Relational—but only the responses from the post-assessment were used to identify a teacher's progress in "understanding how the formal approaches apply to a task and what they mean once executed" (Huey, 2011, p. 68).

4.2. ASSESSMENT OF INFORMAL INFERENTIAL REASONING

Regarding informal inferential reasoning, we also found research that has used the SOLO taxonomy to assess the following elements: inferential processing (Vallecillos & Moreno, 2003), reasoning on inferential tasks, such as the comparison of groups (e.g., Huey, 2011; Nor & Idris, 2010; Watson, 2008), and tasks proposed by Zieffler et al. (2008) to promote IIR in students (Goss, 2014; Jacob, 2013). Additionally, we determined that the AIRS test (Park, 2012) assessed aspects related to reasoning based on informal inference. Vallecillos and Moreno (2003) adapted the SOLO model to evaluate the cognitive development of four constructs—aspects of inferential reasoning that they considered basic to statistical inference thinking in secondary school students: (1) populations, samples, and their relationships, (2) the inferential process, (3) sample size, and (4) types of sampling and biases. By the inferential process, they referred to students' conceptions about the process that make it possible to describe a population based on information from the sample.

Starting with the idea that cognitive development occurs gradually over several stages of schooling, they built a four-level scheme for assessing students' inferential reasoning. First, they administered a 24-item questionnaire to 49 secondary school students and classified the responses according to the four constructs shown in their reasoning. For example, for the inferential process, they classified students' reasoning on four levels: idiosyncratic (if students did not use the information derived from the sample to make statistical inferences, but drew conclusions based on personal or previous ideas), transitional (if they thought that a population could only be described by conducting a census and not by analyzing sample data), quantitative (if they perceived that the inferential process makes it possible to describe the population using the same characteristics as those observed in a sample), and analytical (if they visualized randomness and realized that it is not possible to determine the characteristics of the population precisely using sample information).

Clearly, Vallecillos and Moreno (2003) focused on ascertaining whether students could articulate the uncertainty that is characteristic of the inferential process. They did not expect that students would quantify the uncertainty, so only the informal idea of uncertainty was evaluated. Regarding the other aspects, these authors focused on observing whether students could relate conceptual ideas in different

contexts (e.g., sample and population, sample size, and parameter estimation), and identify the appropriate sampling method for making an inference.

Nor and Idris (2010) and Watson (2008), meanwhile, classified students' IIR level in relation to group comparison informal inferences. Whereas Nor and Idris described the IIR levels of postgraduate students as an example of how IIR could be assessed, Watson studied middle school students (aged 12–13 years). To explore postgraduate students' reasoning, Nor and Idris showed them a comparative box plot that represented data from a random sample. The students first had to indicate whether they could generalize conclusions about the whole population based on that information, and then justify their responses. The difference between these two studies is that Watson's gave participants the freedom to decide how to compare the data that they, or their teacher, collected in their school (e.g., resting heart rate and heart rate during physical activity, arm span), so they could explore the data through various types of graphs and then draw conclusions about populations based on those explorations. In that context, the students could make informal inferences about two groups (e.g., the difference between resting heart rate and heart rate after physical activity, and differences in arm span according to gender or school grade). To do this, the students used dynamic software (TinkerPlots™) during four course sessions.

At the conclusion of each session, the files generated while working with the software were collected and the students were instructed to add text comments on their interpretation of the analyses and informal inferences. The first two sessions focused on the context of heart rate (resting and after physical activity), while in the last two, the context was arm span compared between boys and girls, or between adolescents in different grades. In each session, and for each student, Watson observed whether their comments presented some, or all, of the first eight elements that Pfannkuch (2006b) identified as relevant in group IIR comparisons using box plots: (1) hypothesis generation, (2) summary, (3) shift, (4) signal, (5) spread, (6) sampling, (7) explanatory or context, and (8) individual case. Finally, Watson grouped similar comments by individual students to facilitate classifying IIR levels according to SOLO taxonomy, which allowed detection of the elements of IIR exhibited by students and pinpointed their level of reasoning. IIR levels were thus related to the number of elements that students used in their arguments, and the connections observed among them.

In both studies, the characterization of IIR level in relation to group comparisons focused on analyzing the following aspects: informal inference made by students, distinguishing aspects of data distribution (e.g., center, dispersion, outliers) students used to justify their informal inferences, and observing how students integrated those elements in their arguments. To assess IIR, Watson (2008) and Nor and Idris (2010) coincided in focusing on inference justification, argument coherence, and the integration of distributional aspects. They differ, however, in that Nor and Idris thought that at the highest IIR level (Extended-abstract), the uncertainty associated with the inferential result should be evaluated without a specific request to do so, and aspects that could affect its result should be mentioned (e.g., type of sampling, research design). Watson, in contrast, did not consider evaluating uncertainty, perhaps due to the scope of her study, the difference in school level of participants, and the fact that the Extended-abstract level of SOLO taxonomy was not considered.

In addition to evaluating FIR, Huey (2011) also assessed IIR "by comparing two or more aspects of data when generating an inference without the assistance of a formal algorithm" (p. 3). The pre-assessment, clinical interview, and the post-assessment through which Huey assessed pre-service teachers' IIR included comparing distributions tasks. Huey characterized teachers' responses using the SOLO levels (Pre-structural, Uni-structural, Multi-structural, and Relational) emphasizing only one of the key components of IIR proposed by Zieffler et al. (2008) and Makar and Rubin (2009), namely evidence based on exploratory data analysis. The IIR progress was determined by noticing and incorporating variance or spread concepts in addition to measures of centers to compare data sets; at the highest level (Relational), responses also included the integration of statistical information and context.

Goss (2014) designed an instrument to assess IIR—Assessment of Informal Inferential Reasoning (AIIR)—and then administered it to approximately 900 middle school students. To construct AIIR, Goss used the three types of tasks that Zieffler et al. (2008) recommended to intentionally develop this type of reasoning in students: (1) estimating and drawing a population graph based on the sample data, (2) making inferences about two populations according to sample data, and (3) judging between two competing models or statements. An adaptation of SOLO taxonomy was used to analyze and assess

students' use of IIR in their responses. In a pilot study, Goss decided to use three cycles of the SOLO model, taking the three key IIR principles proposed by Zieffler et al. (2008) as indicators: (1) making judgments and predictions, (2) using and integrating prior knowledge, and (3) articulating arguments based on evidence with a focus on variability and uncertainty. Goss visualized the sophistication of the students' reasoning structure (cognitive network) to determine the cycle to which they belonged. Hence, the first step was to define the three cycles. For example, by focusing on variability expressions that students could use in their generalizations, Goss explained that in

- Cycle 1: Variability was not considered when drawing conclusions.
- Cycle 2: Uncertainty was expressed, but erroneous conclusions were drawn.
- Cycle 3: Uncertainty was expressed, and appropriate conclusions were drawn.

Goss noticed that these descriptors limited observation of how students could use sample variability, so in Goss's final proposal, Goss continued to use three SOLO cycles, but changed the indicators to evaluate (1) use of variability, (2) use of context, and (3) certainty and argumentation. As in the pilot study, Goss first ascertained the cycle to which students belonged, and then assessed the number and use of structures to determine each student's level within that cycle. Goss added a level of Pre-structural reasoning prior to the three cycles. Thus, for each of the three types of tasks on the AIIR test, Goss visualized 10 levels of reasoning, as follows: Pre-structural, U_1, M_1, R_1, U_2, M_2, R_2, U_3, M_3 and R_3 (U = Uni-structural, M = Multi-structural, R = Relational; the numbers correspond to the cycle number). In this way, Goss described how students' IIR developed.

In general, Goss (2014) focused on determining the level at which students used context knowledge or context information to make an informal inference, the language with which they expressed their inferences (deterministic or probabilistic), their use of conceptual ideas of variability to make an inference, and whether they could integrate different aspects of distribution (e.g., center, spread, outliers, shape).

The aspects of IIR evaluated in AIIR (Goss, 2014) were the language of uncertainty, the concept of sampling variability, the prediction of characteristics of a population based on a sample and whether two populations were judged to be similar or different based on samples, and the use of context information. Park (2012), in contrast, focused on assessing whether students were able to reason about informal conceptual ideas (e.g., uncertainty, unusualness, inference), use fundamental concepts or properties (e.g., aggregate property, sampling variability) to make an informal inference, and articulate the relationship between sample size and the shape of the distribution of sample statistics.

Jacob (2013) also analyzed 136 students' IIR (Grades 11 and 12) development at two different schools. At the beginning of an introductory statistics course, Jacob administered a pre-test assessment containing informal statistical inference questions. All students then took part in three informal statistical inference classroom activities during a school year. These activities addressed informal statistical inference about (1) comparison of data distributions, (2) sampling and probability, and (3) sampling distribution. At the end of each of the three informal classroom inference activities, four task-based interviews were conducted with seven pairs of students. Three included informal inferential tasks related to the same topics addressed in mentioned activities. At the end of the course, all students answered a post-test assessment that contained 12 questions on informal and 10 on formal statistical inference.

In investigating students' IIR development, Jacob (2013) observed a substantial gain in their responses from the pre- to post-test questions in drawing an informal conclusion using important aspects of data sets (center, variability, and shape), based on sampling and estimating a probability and by comparing a sample statistic to its related sampling distribution. Indeed, Jacob reported that the students' responses for the three classroom informal activities and the three informal task-based interviews "were analyzed to determine if they were able to make informal inferences that included the three main principles of informal statistical inference described by Makar and Rubin (2009)" (p. 47).

4.3. SIMILARITIES AND DIFFERENCES IN THE ASSESSMENT OF INFORMAL AND FORMAL INFERENTIAL REASONING

In this section we present a summary of the differences and similarities in the assessment of IIR and FIR (see Table 5) and in the focus of their assessment (see Table 6). We can observe that IIR was evaluated using more qualitative approaches (i.e., the SOLO taxonomy), than quantitative approaches. The aspects evaluated in both approaches (FIR and IIR) are coherent with their conceptions. On the one hand, FIR statistical ideas evaluated are related to concepts of statistical inference whereas those of IIR are related to descriptive statistics; on the other hand, the focus of the evaluation in FIR involves more formal knowledge and IIR a more exploratory data analysis. The role of context was considered in both approaches; it seems that in the FIR assessment, the context was more considered for interpretation of results of statistical tests, and in the IIR assessment to make the inference.

Table 5. Similarities and differences in the assessment of IIR and FIR

Assessment of IIR and FIR
Similarities
1) The similar types of instruments used to assess FIR and IIR were validated tests (RPASS-7, AIRS and AIIR), task-based interviews, questionnaires, and post-tests.
2) The use of SOLO taxonomy to characterize students' levels of reasoning (formal or informal) according to their responses.
3) Ideas of population, sample size, sampling, and *p*-value were considered from the point of view of the two approaches (IIR and FIR).
Differences
1) Other types of instruments were used only to assess IIR: classroom activities, pre-test, and pre-clinical interview.
2) Although SOLO was used to assess FIR (we found two studies), it was used more to assess IIR.
3) In relation to the use of validated tests, only the responses of the AIIR test administration were analyzed using SOLO taxonomy; RPASS-7 and AIRS responses were evaluated as correct/incorrect or valid/invalid.

Table 6. Similarities and differences in the focus of the assessment of IIR and FIR

Focus of Assessment of IIR and FIR
Similarities
1) The common focus of assessment in both approaches was the conceptual knowledge, integration, and use of statistical ideas.
2) The knowledge context of the inferential problem was considered in the interpretations of the conclusion.
Differences
1) FIR assessment focused on definitions of formal ideas or concepts of inference (*p*-value, confidence interval, and significance level), interpretations of results of statistical tests (hypothesis test, confidence intervals, statistical significance, and *p*-value), the choice of the appropriate statistics in hypothesis testing, necessary conditions of a study design, meaning of confidence interval, the global logic of the hypothesis testing processes, the formulation of a hypothesis, the relationship between sample size and width of the confidence interval and relationship of *p*-value to sample size and statistical significance, identifying the difference between statistical and practical significance given the context of the inferential situation, and the correct execution of inference methods.
4) IIR assessment focused on generalizing from a sample to a population, making judgments and predictions, using and integrating prior knowledge, articulating arguments, use of context to make an inference, use of properties of aggregates, relating sample size and the distribution of sample statistics, estimating and drawing a population graph based on the sample data, measuring uncertainty, use of language to express inferences (deterministic or probabilistic), reasoning about conceptual ideas, estimating a probability, and integration of statistical information and context.

5. ASSESSING THE RELATIONSHIP BETWEEN IIR AND FIR

In this section, we discuss three studies that analyzed the relationship between IIR and FIR: the doctoral dissertations by Huey (2011), Park (2012), and Jacob (2013). Huey (2011) and Jacob (2013) used pre-test and post-test evaluations, as well as interviews. In Huey's research, the reasoning levels of the SOLO cognitive model were used to characterize the responses of those evaluations, whereas Jacob used both qualitative and quantitative techniques. Park (2012), in contrast, analyzed whether the AIRS test could evaluate students' inferential reasoning by distinguishing between IIR and FIR.

5.1. ASSESSING THE ASSOCIATION BETWEEN DOMINANT LEVELS OF IIR AND FIR

The investigation by Huey (2011) addressed the study of relationships between the change in informal and formal reasoning using correlation techniques to measure the association. Because participants (preservice teachers) would provided formal or informal responses in the post-assessment tasks, responses from the post-assessment were used to measure the association (189 informal responses compared to 49 formal responses). For each teacher, Huey observed whether or not the teacher's reasoning level on the formal task was the same as the dominant IIR level. If it was dominant, it was coded as concordant; if not, it was coded as discordant.

Huey (2011) found a relationship between the informal and formal levels for most teachers, as 80% of the 49 formal responses were concordant. According to these results, she indicated that there seems to be a relatively strong relationship between the inferential reasoning ability shown by these teachers on informal inferential tasks and their ability to execute formal ones. That is, "the preservice teachers tended to reason at the similar levels for both formal and informal approaches on the post-assessment" (p. 137). In this sense, the association consisted of comparing between informal and formal responses on equivalent levels of reasoning.

One of the limitations Huey (2011) identified is that her work did not assess *adapted reasoning*. By adapted reasoning, she referred to analyzing teachers' reasoning in terms of their logic, reflections, and explanations in relation to the justification of the inference. Huey therefore suggested that more research was necessary, as assessing this type of reasoning may improve our comprehension of how relations among conceptual inference ideas are being understood.

5.2. ASSESSING THE ASSOCIATION BETWEEN IIR AND FIR SCORES

Regarding the association between students' IIR and FIR, Jacob (2013) was interested in analyzing measuring the correlation of the scores for two groups of participants: the students who showed substantial improvement in IIR by course end, and the ones who showed great skill in IIR at the beginning and end of the course. On the basis of the results obtained from correlating scores from these two outstanding groups, Jacob concluded that

- there was no statistically significant positive correlation between the IIR and FIR scores of the student group that showed substantial improvement in IIR at the end of the course.
- a statistically significant positive correlation existed between the IIR and FIR scores of the group with high IIR ability at the beginning and end of the course, due to the strong association that those students showed between sub-scores when making inferences from a population based on sample data, and the scores attained in formal statistical inference.

According to Jacob's (2013) quantitative results, the significant positive correlation observed in students with high IIR abilities at the beginning and end of the course indicated that those who began the course with good IIR abilities were skilled reasoners in both types of inference (informal and formal). In contrast, although the students who were not good at IIR at the beginning of the course did improve their IIR abilities, they did not necessarily advance in FIR. Using qualitative analysis, Jacob revealed various characteristics of students' IIR and FIR. For example, Jacob pointed out that the conclusions from students' IIR on distribution comparisons was based mainly on the use of means or medians. Although they recognized variability in a distribution, they did not use it to make their inferences. Also, Jacob observed that students tended to use more procedural knowledge over conceptual knowledge in their FIR (i.e., confidence interval construction and hypothesis testing procedures). Finally, Jacob recognized that the results cannot be generalized to other high school

populations because the research was limited to only local two institutions. Moreover, Jacob revealed that the students were not selected randomly for the interviews, and that during classroom observation Jacob focused only on the students' work and did not consider that it might be important to observe the teacher as well.

Returning to the AIRS test, we found that Park (2012) analyzed the dimensionality of the response data gathered in order to determine whether inferential reasoning in statistics was unidimensional or multidimensional. The item responses were broken down into informal statistical inferences (ISI) or formal statistical inferences (FSI). Confirmatory factor analysis (CFA) was applied, and two factor models (one unidimensional, the other a two-factor model) were examined and compared. The two-factor model consisted of one factor based on the items that assessed ISI and a second factor based on items that measured FSI. The results of CFA did not support the hypothesized structure of two domains; rather, they indicated that the AIRS responses were unidimensional with a high correlation between IIR and FIR. According to Park, this could be due to an unclear distinction in the literature between informal and formal inference or because students might use informal or formal inference indistinctly. Park further explained that the unidimensionality of the AIRS test suggested that

- "a student who understands the ideas in FSI probably (1) uses FSI when it is required; (2) uses the ideas in FSI when only ISI is needed; or (3) uses both ideas in ISI and FSI when either are required.
- Considering that ISI is foundational to FSI, students with a good understanding of FSI might have a good understanding of ISI, and it may be that those who do not develop a good understanding of ISI have difficulty with developing FSI."
  (Park, 2012, p. 163)

Finally, Park's (2012) explanation coincides with Jacob's (2013) results in that it seems necessary to have a good understanding of IIR to develop a good understanding of FIR.

6. SUMMARY

According to our review, the distinction between FIR and IIR seems to lie in the types of methods, procedures, or formulas used to solve inferential tasks or problems and in the interpretation of their results. In IIR, emphasis is placed on exploratory analysis of data, for example identifying patterns in graphic representations using more visual than numerical characteristics, and estimating inference uncertainty using empirical sampling distributions. FIR, in contrast, refers to the use of formal inference methods (e.g., hypothesis testing, confidence intervals).

Regarding assessments of inferential reasoning, we were able to identify that some researchers have designed and validated instruments (e.g., RPASS, AIRS, AIIR) that aim to assess students' inferential reasoning—whether formal, informal, or both. Other scholars have assessed this reasoning through questionnaires, classroom activities, or pre-tests and post-tests, and then analyzed students' responses as a means of characterizing them in terms of levels of reasoning, according to the SOLO taxonomy (e.g., Goss, 2014; Huey, 2011; Inzunsa & Jiménez, 2013; Nor & Idris, 2010; Vallecillos & Moreno, 2003; Watson, 2008). We observed that those authors agreed on definitions of the levels of the SOLO taxonomy according to the skills students applied to solve a specific inferential task, or diverse types of inferential tasks. These skills are related to the use of procedures and the conceptual understanding and integration of statistical ideas involved in the task. The types of tasks used to evaluate inferential reasoning are associated with the broad concepts that underlie statistical inference (e.g., *p*-value, significance tests, confidence intervals, sampling, sampling distribution, distribution, sampling variability, comparison of groups, and probability). Goss (2014) and Huey (2011) evaluated the use of context to provide a reasonable inference for a particular problem. They consider that in the high level of inferential reasoning students have to consider context and statistics.

The few studies that have assessed a possible relationship between IIR and FIR have used both qualitative and quantitative techniques. In Huey's (2011) study, these two types of reasoning were characterized in terms of SOLO taxonomy levels. Huey assessed FIR according to the selection of the hypotheses tested or formal inference methods, the correct execution of those methods, and a reasonable interpretation of results. Like Huey, Jacob's (2013) approach to assessing FIR was based on the results of interpretations of hypothesis testing and estimations of the confidence interval for population

parameters. Huey assessed IIR in terms of exploratory data analysis and the use of context, whereas Jacob focused on evaluating IIR when applied to comparisons of data distributions, sampling and probability, and sampling distribution. The relationship that Huey detected was based on the concordance between the IIR and FIR levels determined. Huey's results showed that the reasoning levels of almost all participants on a formal inference task were similar to their reasoning levels on the equivalent informal task. In contrast, Jacob showed that an association between IIR and FIR can arise once an individual develops a certain ability to reason informally. That is, someone who shows a great ability to reason informally at the beginning and end of a course ends up also being able to reason formally, whereas people who do not begin with good IIR do not necessarily develop abilities to reason formally, although they may improve their IIR ability by course end. This assumes, of course, that achieving good FIR requires certain prior skills in IIR, which will continue to develop during statistics courses.

Finally, the AIRS test analysis revealed that there is no distinction between ISI and FSI. The result "implies that students might use both informal and formal methods of statistical inference even when they do not need to use formal statistical ideas" (Park, 2012, p. 164). On the one hand, Park's and Jacob's studies confirm what other authors have pointed out in that IIR may provide a pathway to FIR. On the other hand, the studies of Huey (2011), Park (2012), and Jacob (2013) did not inform the question of how IIR evolves into FIR, so it may be a current challenge in research on inferential reasoning.

7. DISCUSSION AND DIRECTIONS FOR FUTURE RESEARCH

Based on our review, we now underscore important aspects of research on inferential reasoning that we believe have great potential for the future. More study is required to distinguish clearly the characteristics, or components, of FIR from those of IIR. According to Park (2013), the term FSI seems to be used interchangeably to refer to formal reasoning about statistical inference, whereas ISI seems to be used to refer to informal reasoning. Hence, FIR is associated with the use and understanding of concepts involved in formal statistical tests (e.g., *p*-value, statistical significance, hypothesis testing, and confidence intervals); conversely, those concepts do not arrise in IIR.

Differentiating IIR from FIR only by the informality or formality of the statistical methods used to make the inference, however, can create a misleading image of a fragmented inferential reasoning, and this may be one of the reasons to be interested in how to help students move from the informal to the formal. On the one hand, inferential reasoning—formal or informal—is complex, because it integrates various statistical concepts (Chance, delMas, & Garfield, 2004); on the other, it requires information about the context of the problem, including the techniques, norms, and proposals used in the field of science to which the problem under study pertains (see Bakker et al., 2008). In this sense, Bakker et al. (2008) distinguish *statistical* and *contextual* reasons to refer to the *space of reason* as the holistic approach to characterizing statistical inference.

The foregoing discussion suggests that inferential reasoning requires both statistical knowledge (statistical reasons) and knowledge of the disciplinary field on which the inferential problem is based (contextual reasons), similar to the continual interaction between the context and statistical spheres that are required for statistical thinking (Wild & Pfannkuch, 1999). In our review, we observed that the role of context is considered in different ways to assess students' inferential reasoning, whether informal or formal. For instance, context is contemplated in the problems, tasks, or items used in the evaluation, whereas in the SOLO levels of reasoning, when students made an inference based on data and context, they tended to place their responses at a high level of reasoning (e.g., Goss, 2014; Huey, 2011). Also, context is assessed when students take it into account to choose and formulate the correct hypothesis test, and to make appropriate interpretations of statistical results (e.g., Huey, 2011; Inzunsa & Jiménez, 2013; Lane-Getaz, 2013; Park, 2012), as well as to distinguish practical significance given the context of the inferential situation (Lane-Getaz, 2013).

We found that more attention has been paid to assessing reasoning in terms of how students use or integrate statistical ideas or concepts—either informally or formally—in making an inference, and the context has been taken into account to a lesser extent. If inferential reasoning is comparable to the argumentation process, then there is a need to redefine our conception of it, because argumentation is "the whole activity of making claims, challenging them, backing them up by producing reasons,

criticizing those reasons, rebutting those criticisms, and so on" (Toulmin, Rieke, & Janik, 1984, p. 14). Thus, when individuals argue to support a claim, or to make or justify a decision, they need to present reasons. In this setting, what role do contextual and statistical reasons play in inferential reasoning? From our perspective, both reasons function as the evidence upon which the statistical inference is based. On the one hand, statistical reasons use data to draw conclusions on the basis of the meaning of statistical results; on the other, contextual reasons may serve to connect the statistical reasons to the conclusions: contextual reasons examine under what circumstances the given statistical results support the conclusion. This is because, according to Bakker and Derry (2011), "what counts as valid reasoning, adequate judgment or correct application of concepts depends on the norms being used in a particular practice" (p. 12).

Redefining the conception of inferential statistical reasoning—that is, beyond classifying it as informal or formal—requires pondering the use of different theories to analyze students' inferential reasoning (see the proposal by Gómez-Blancarte & Tobías-Lara, 2018). Moreover, it may make it unnecessary to inquire into the transition between informal and formal inferential reasoning. Instead, we may ask about the central activity of reasoning, that is, focus on evaluating students' inferential reasoning in terms of the articulation of the statistical and contextual reasons they present to support their inferences.

This article has provided an overview of research on assessments of IIR and FIR. We trust that our discussion of how these types of reasoning have been understood and how their assessment has been addressed will provide points for reflection as we rethink the conception of inferential reasoning and identify the aspects upon which attention needs to be focused.

REFERENCES

Bakker, A., & Derry, J. (2011). Lessons from inferentialism for statistics education. *Mathematical Thinking and Learning*, *13*(1–2), 5–26.
[Online: https://doi.org/10.1080/10986065.2011.538293]

Bakker, A., Kent, P., Derry, J., Noss, R., & Hoyles, C. (2008). Statistical inference at work: Statistical process control as an example. *Statistics Education Research Journal*, *7*(2), 130–145.
[Online: http://iase-web.org/documents/SERJ/SERJ7(2)_Bakker.pdf]

Batanero, C. (2000). Controversies around the role of statistical tests in experimental research. *Mathematical Thinking and Learning*, *2*(1–2), 75–98.

Batanero (2011, June). Del análisis de datos a la inferencia: Reflexiones sobre la formación del razonamiento estadístico [From data analysis to inference: Reflections about the formation of statistical reasoning]. Paper presented at *XIII Conferencia Interamericana de Educación Matemática* (pp. 277–291). Recife, Brazil.
[Online paper: http://revistas.ucr.ac.cr/index.php/cifem/article/view/14732]

Ben-Zvi, D. (2006). Scaffolding students' informal inference and argumentation. In A. Rossman & B. Chance (Eds.), *Working cooperatively in statistics education. Proceedings of the Seventh International Conference on Teaching Statistics*, Salvador, Brazil. Voorburg, The Netherlands: International Statistical Institute.
[Online: https://iase-web.org/documents/papers/icots7/2D1_BENZ.pdf]

Ben-Zvi, D., Aridor, K., Makar, K., & Bakker, A. (2012). Students' emergent articulations of uncertainty while making informal statistical inferences. *ZDM–International Journal on Mathematics Education*, *44*(7), 913–925.
[Online: https://doi.org/10.1007/s11858-012-0420-3]

Ben-Zvi, D., Gil, E., & Apel, N. (2007). What is hidden beyond the data? Young students reason and argue about some wider universe. In D. Pratt & J. Ainley (Eds.), *Reasoning about statistical inference: Innovative ways of connecting chance and data. Proceedings of the Fifth International Research Forum on Statistical Reasoning, Thinking and Literacy (SRTL-5)* (pp. 1–26). Warwick, England: University of Warwick.
[Online: https://blogs.uni-paderborn.de/srtl/srtl5/]

Chance, B., delMas, R., & Garfield, J. (2004). Reasoning about sampling distributions. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 295–323). Minnesota: Springer Science.

Cobb, G. W., & Moore, D. S. (1997). Mathematics, statistics, and teaching. *The American Mathematical Monthly*, *104*(9), 801–823.

García, V. N., & Sánchez, E. A. (2014). Razonamiento inferencial informal: El caso de la prueba de significación con estudiantes de bachillerato [Informal inferential reasoning: The case of the significance test with high school students]. In M. T. González, M. Codes, D. Arnau, & T. Ortega (Eds.), *Investigación en Educación Matemática XVIII* (pp. 345–354). Salamanca: SEIEM. [Online: www.seiem.es/pub/actas/index.shtml]

Garfield, J., & Ben-Zvi, D. (2008). Learning to reason about statistical inference. In J. Garfield & D. Ben-Zvi (Eds.), *Developing students' statistical reasoning: Connecting research and teaching practice* (pp. 351–389). Minnesota: Springer Science.

Garfield, J., delMas, B., & Zieffler, A. (2012). Developing statistical modelers and thinkers in an introductory, tertiary-level statistical course. *ZDM–International Journal on Mathematics Education*, *44*(7), 883–898. [Online: https://doi.org/10.1007/s11858-012-0447-5]

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. [Online: https://doi.org/10.1007/s10649-014-9541-7]

Garfield, J., & Zieffler, A. (2011). Informal and formal statistical inference?: New questions raised. Discussion on the paper by Wild, Pfannkuch, Regan, and Horton *Journal of the Royal Statistical Society*, *174*(2), 280. [Online: https://doi.org/10.1111/j.1467-985X.2010.00678.x]

Gil, E., & Ben-Zvi, D. (2014). Long-term impact on students' informal inferential reasoning. In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in Statistics Education*. *Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS-9)*, Flagstaff, Arizona, USA. Voorburg, The Netherlands: International Statistical Institute. [Online: https://iase-web.org/icots/9/proceedings/pdfs/ICOTS9_8D1_GIL.pdf]

Gómez-Blancarte, A., & Tobías-Lara, M. G. (2018). Using the Toulmin model of argumentation to validate students' inferential reasoning. In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking Back, Looking Forward. Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS-10)*, Kyoto, Japan. Voorburg, The Netherlands: International Statistical Institute. [Online: http://iase-web.org/icots/10/proceedings/pdfs/ICOTS10_8C1.pdf]

Goss, J. M. (2014). *A method for assessing and describing the informal inferential reasoning of middle school students* (Unpublished doctoral dissertation). Western Michigan University, Kalamazoo, MI. [Online: http://scholarworks.wmich.edu/cgi/viewcontent.cgi?article=1280&context=dissertations]

Huey, M. E. (2011). *Characterizing middle and secondary preservice teachers' change in inferential reasoning* (Unpublished doctoral dissertation). University of Missouri-Columbia, Columbia, MO. [Online: http://iase-web.org/documents/dissertations/11.MaryannEHuey.Dissertation.pdf]

Inzunsa, S., & Jiménez, J. (2013). Caracterización del razonamiento estadístico de estudiantes universitarios acerca de las pruebas de hipótesis [The characteristics of college students' statistical reasoning on hypothesis testing], *Revista Latinoamericana de Investigación en Matemática Educativa*, *16*(2), 179–211. [Online: http://doi.org/10.12802/relime.13.1622]

Jacob, B. L. (2013). *The development of introductory statistics students' informal inferential reasoning and its relationship to formal inferential reasoning* (Unpublished doctoral dissertation). Syracuse University, Syracuse, NY. [Online: https://surface.syr.edu/tl_etd/245/]

Jacob, B. L., & Doerr, H. M. (2013). Students' informal inferential reasoning when working with the sampling distribution. In B. Ubuz, C. Haser, & M. A. Mariotti (Eds.), *Proceedings of the Eighth Congress of the European Society for Research in Mathematics Education* (pp. 829–839). Ankara, Turkey: Middle East Technical University and ERME.

Johnson, R. (2011). Statistical inference. In M. Lovric (Ed.), *International Encyclopedia of Statistical Science* (pp. 1418–1420). Berlin: Springer-Verlag.
[Online: https://doi.org/10.1007/978-3-642-04898-2]

Lane-Getaz, S. J. (2007). Toward the development and validation of the reasoning about *p*-values and statistical significance scale. In L. Weldon, B. Phillips, & T. Shea (Eds.), *Proceedings of the ISI/IASE Satellite Conference on Assessing Student Learning in Statistics.* Guimarães, Portugal. Voorburg, The Netherlands: International Statistical Institute.
[Online: http://iase-web.org/Conference_Proceedings.php?p=Assess_Stud_Learn_2007]

Lane-Getaz, S. J. (2013). Development of a reliable measure of students' inferential reasoning ability. *Statistics Education Research Journal*, *12*(1), 20–47.
[Online: http://iase-web.org/documents/SERJ/SERJ12(1)_LaneGetaz.pdf]

Makar, K. (2013). Predict! Teaching statistics using informal statistical inference. *Australian Mathematics Teacher*, *69*(4), 34–41.

Makar, K., Bakker, A., & Ben-Zvi, D. (2011). The reasoning behind informal statistical inference. *Mathematical Thinking and Learning, 13*(1), 152–173.
[Online: https://doi.org/10.1080/10986065.2011.538301]

Makar, K., & Rubin, A. (2009). A framework for thinking about informal statistical inference. *Statistics Education Research Journal*, *8*(1), 82–105.
[Online: http://iase-web.org/documents/SERJ/SERJ8(1)_Makar_Rubin.pdf]

Makar, K., & Rubin, A. (2014). Informal statistical inference revisited. In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in Statistics Education*. *Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS-9),* Flagstaff, Arizona, USA. Voorburg, The Netherlands: International Statistical Institute.
[Online: http://icots.info/9/proceedings/pdfs/ICOTS9_8C1_MAKAR.pdf]

Moore, D. S. (1997). New pedagogy and new content: The case of statistics. *International Statistical Review, 65*(2), 123−137. doi: 10.2307/1403333
[Online: https://doi.org/10.2307/1403333]

Nor, N. M., & Idris, N. (2010). Assessing students' informal inferential reasoning using SOLO taxonomy based framework. *Procedia Social and Behavioral Sciences*, *2*(2), 4805–4809.
[Online: https://doi.org/10.1016/j.sbspro.2010.03.774]

Park, J. (2012). *Developing and validating an instrument to measure college students' inferential reasoning in statistics: An argument-based approach to validation* (Unpublished doctoral dissertation). University of Minnesota, Minneapolis, MN.
[Online: http://www.stat.auckland.ac.nz/~iase/publications/dissertations/12.Park.pdf]

Park, J. (2013). Designing an assessment to measure students' inferential reasoning in statistics: The first study, development of a test blueprint. *Research in Mathematical Education* (Korean Society of Mathematical Education), *17*(4), 243–266.
[Online: http://doi.org/10.7468/jksmed.2013.17.4.243]

Pfannkuch, M. (2006a). Informal inferential reasoning. In A. Rossman & B. Chance (Eds.), *Working cooperatively in statistics education. Proceedings of the Seventh International Conference on Teaching Statistics*, Salvador, Brazil. Voorburg, The Netherlands: International Statistical Institute.
[Online: https://iase-web.org/documents/papers/icots7/6A2_PFAN.pdf]

Pfannkuch, M. (2006b). Comparing box plot distributions: A teacher's reasoning. *Statistics Education Research Journal*, *5*(2), 27–45.
[Online: http://iase-web.org/documents/SERJ/SERJ5(2)_Pfannkuch.pdf]

Pfannkuch, M. (2011). The role of context in developing informal statistical inferential reasoning: A classroom study. *Mathematical Thinking and Learning, 13*(1–2), 27–46.
[Online: https://doi.org/10.1080/10986065.2011.538302]

Pratt, D., & Ainley, J. (2008). Introducing the special issue on informal inferential reasoning. *Statistics Education Research Journal*, *7*(2), 3–4.
[Online: http://iase-web.org/documents/SERJ/SERJ7(2)_Pratt_Ainley.pdf]

Reading, C. (2007, Aug.). Cognitive development of reasoning about inference. Discussant reaction presented at the *Fifth International Forum for Research on Statistical Reasoning, Thinking and Literacy (SRTL-5)*, University of Warwick, England.

Rossman, A. J. (2008). Reasoning about informal statistical inference: One statistician's view. *Statistics Education Research Journal*, *7*(2), 5–19.
[Online: http:/iase-web.org/documents/SERJ/SERJ7(2)_Rossman.pdf]

Rossman, A. J., & Chance, B. L. (1999). Teaching the reasoning of statistical inference: A "top ten" list. *The College Mathematics Journal*, *30*(4), 297–305.

Teng, C. H. (2016). Informal inferential reasoning, graphical representations and the Singapore primary statistics curriculum. *The Mathematics Educator*, *16*(2), 83–108.

Toulmin, S., Rieke, R., & Janik, A. (1984). *An introduction to reasoning* (2nd ed.). New York: Macmillan Publishing Company.

Vallecillos, A., & Batanero, C. (1997). Aprendizaje y enseñanza del contraste de hipótesis: Concepciones y errores [Learning and teaching of hypotheses' contrasts: Conceptions and mistakes]. *Ensenanza de las Ciencias*, *15*, 189–197.

Vallecillos, A., & Moreno, A. J. (2003). Esquema para la instrucción y evaluación del razonamiento en estadística [Scheme for the instruction and evaluation of reasoning in basic inferential statistics]. *Revista de Educación y Pedagogía*, *15*(35), 71–81.

Watson, J. M. (2008). Exploring beginning inference with novice grade 7 students. *Statistics Education Research Journal*, *7*(2), 59–82.
[Online: http://iase-web.org/documents/SERJ/SERJ7(2)_Watson.pdf]

Watson, J. M., & Moritz, J. B. (1998). The beginning of statistical inference: Comparing two data sets. *Educational Studies in Mathematics*, 37(2), 145–168.
[Online: https://doi.org/10.1023/A:1003594832397]

Weinberg, A., Wiesner, E., & Pfaff, T. J. (2010). Using informal inferential reasoning to develop formal concepts: Analyzing an activity. *Journal of Statistics Education*, *18*(2), 1–23.
[Online: https://doi.org/10.1080/10691898.2010.11889494]

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review*, *67*(3), 223–265. doi: 10.1111/j.1751-5823.1999.tb00442.x
[Online: https://doi.org/10.1111/j.1751-5823.1999.tb00442.x]

Wild, C. J., Pfannkuch, M., Regan, M., & Parsonage, R. (2017). Accessible conceptions of statistical inference: Pulling ourselves up by the bootstraps. *International Statistical Review*, *85*(1), 84–107.
[Online: https://doi.org/10.1111/insr.12117]

Zieffler, A., Garfield, J., delMas, R., & Reading, C. (2008). A framework to support research on informal inferential reasoning. *Statistics Education Research Journal*, *7*(2), 40–58.
[Online: http://iase-web.org/documents/SERJ/SERJ7(2)_Zieffler.pdf]

ANA LUISA GÓMEZ-BLANCARTE
Instituto Politécnico Nacional, CICATA-Legaria
Calzada Legaria No. 694, Col. Irrigación
C.P. 11500, Del. Miguel Hidalgo
Ciudad de México, MÉXICO