Tertiary Students' Understanding of Sampling Distribution

Adeola Ajao
University of Tasmania
adeola.ajao@utas.edu.au

Noleine Fitzallen
University of New South Wales
n.fitzallen@unsw.edu.au

Helen Chick
University of Tasmania
helen.chick@utas.edu.au

Greg Oates
University of Tasmania
greg.oates@utas.edu.au

In this paper, the SOLO taxonomy is used to identify different levels of student understanding of the statistical concepts associated with sampling distribution. This study was part of a research project investigating students' conceptual understanding of concepts of hypothesis testing taught with the support of simulation learning activities. The study involved eight students enrolled in a first-year tertiary introductory statistics unit and the examination of their written responses to three questions about sampling distribution concepts. The SOLO taxonomy categorisations revealed that some students had only pre- and unistructural understanding of sampling distribution, and none providing responses at the extended abstract level.

The teaching of statistics has developed and evolved both in terms of curricula and pedagogical practices, yet statistics is still a difficult subject to teach and learn (Horton, 2015). Over time, researchers have investigated what makes the subject of statistics difficult. For example, Watts (1991) suggested that maybe the terms used in statistics—such as "a mean", "variance", "probability," or "a random variable"—together with the abstract nature of the associated concepts are the crux of the difficulties students encounter. Other factors associated with students' difficulties include changes in curriculum (Duckworth & Stephenson, 2002), understanding ideas like resampling and random distribution (Larwin & Larwin, 2011), debate about probability as a separate discipline from statistics (Zieffler et al., 2018), inability to meaningfully interpret problems in statistics (Reaburn, 2014), understanding statistical inference (Chance et al., 2022; Kula & Koçer, 2020), and the lack of individual experience with data when teaching certain topics like sampling distribution (Sued & Valdora, 2021).

The ideas of population and sample are two key concepts in inferential statistics. Students must first understand that the sample they are using is merely one of a huge number of samples that may be taken from the population, and next, to draw conclusions, that the distribution of the means of these samples must be known or modelled. The "sampling distribution" is a crucial idea in statistical inference; the outcome of repeatedly drawing samples from a population of a fixed size, calculating the sample statistic's value (invariably, the mean) for each sample, and then forming a distribution of those values (see, e.g., Bruce, 2014). However, this concept is often poorly understood (Ozmen & Guven, 2019; Sued & Valdora, 2021; Watkins et al., 2014). This report focuses on the learning of sampling distribution.

Why is the concept of sampling distribution hard to teach and learn? Sampling distribution is a multi-faceted abstract concept compared to elementary statistical ideas (Kula & Koçer, 2020; Watts, 1991). In addition, the sampling distribution is more an hypothetical distribution than it is an experiential distribution (Watkins et al., 2014). That is, when students learn sampling distribution, they cannot experience building a full sampling distribution, but instead must abstract it. As well, students commonly develop misconceptions of the complex terminology that is essential to understanding sampling distribution. These misconceptions involve the sample mean, effect of sample size, variance in sample means, standard deviation (Watkins et al., 2014), standard error of

(2023). In B. Reid-O'Connor, E. Prieto-Rodriguez, K. Holmes, & A. Hughes (Eds.), *Weaving mathematics education research from all perspectives*. *Proceedings of the 45th annual conference of the Mathematics Education Research Group of Australasia* (pp. 73–80). Newcastle: MERGA.

the mean, and improper use of probability language (Chance et al., 2004). Student misconceptions about sampling distribution include:

- The sampling distribution should look like the population distribution (for $n > 1$),
- Sampling distributions for small samples and large samples have the same variability,
- Sampling distributions for large samples have more variability,
- A sampling distribution is not a distribution of sample statistics,
- One sample (of real data) is confused with all possible samples (in distribution) or potential samples,
- The law of large numbers (larger samples better represent a population) is confused with the central limit theorem (distributions of means from large samples tend to form a normal distribution), and
- The mean of a positive skewed distribution will be greater than the mean of the sampling distribution for large samples taken from this population (Ben-Zvi, 2004).

To understand sampling distribution, it is necessary to conceptualise sampling from a population, sampling variability, effect of sample size, long run frequency, that is, knowing that the process of random selection causes variability in outcomes, but a stable distribution is achieved over a long run (Pfannkuch et al., 2012), and an understanding of the relationship between population parameters and sample means(Ozmen & Guven, 2019).

One teaching strategy employed in tertiary statistics education is the use of simulations as a means of depicting real life situations. Statistical simulations allow multiple retrials of a given sample to generate a larger data size (Blejec, 2003), and are known as a computer simulation when a computer is used as a means of generating virtual data quickly. Simulations are believed to be effective because it takes advantage of the dual relationship between a distribution and a sample from that distribution. Learners can control the simulation, allowing them to discover the effects of changing the sample size by generating various sample sizes (Hesterberg, 2015).

Although, the teaching of sampling distribution has been researched extensively (Garfield et al., 2015; Kula & Koçer, 2020; Ozmen & Guven, 2019; Watkins et al., 2014), the research focus has been mostly on statistical inference. For example, Chance et al. (2022), Makar and Rubin (2018), Morris et al. (2019), Rossman and Chance (2014), Sigal and Chalmers (2016) all explored tertiary student understanding of sampling distribution within the context of statistical inference but with using simulation. The various principles related to sample distribution are what make it complex. Using the SOLO taxonomy (further discussed in the analysis section), Watson (2004) studied pupils at various stages of reasoning about samples. Using the SOLO taxonomy, an analysis of variation—another concept connected to sampling distribution—showed different levels of knowledge development in a primary six class (Watson et al., 2022). The SOLO taxonomy will be used here to analyse students' understanding of sample distribution at various levels as ideas connected to sampling distribution are investigated. Hence, the research question addressed in this paper is:

- What does a SOLO taxonomy categorisation tell us about students' understanding of sampling distribution?

Research Methods

The research reported in this paper adopted a pragmatist paradigm (Mackenzie & Knipe, 2006) to explore tertiary students' understanding of sampling distribution. The methodology was chosen to capture specifically the students' level of understanding of sampling distribution demonstrated using designated statistical simulations. The research was conducted in a first-year undergraduate statistics unit over three semesters at a regional Australian university. The unit, *Data Handling and Statistics 1,* was taken by students from courses that require a foundational statistics unit and was offered over a period of 12 weeks in both Semester 1 and 2 of the academic year. Students in this

study participated in the unit by attending weekly 1-hour lectures and 2-hour tutorial sessions. Assessment of the unit comprised written responses to quiz questions, completion of projects, and a final examination.

Four simulation activities that focused on different statistical topics were implemented as a pedagogical intervention in lectures. The focus of this report is the extension tasks that followed one of these simulation activities, which focussed on sampling distribution. The learning intentions of this activity and the extension tasks were to understand sampling distribution, simulation, sampling bias, and randomness. The first author participated as an observer during the implementation of the activity. The data collected for this study comes from the eight participants who were enrolled in the unit, gave informed consent to participate in the research, and provided written responses to the extension tasks. These were completed in their own time after the lecture that included the simulation activity. Participants also engaged in after-class activities, one of which is reported and discussed in this paper.

Task

Scenarios used in the simulation activity and the questions used in the extension tasks afterwards were sourced from a moderated statistics blog called, *Ask Good Questions: A Blog About Teaching Introductory Statistics* (https://askgoodquestions.blog/). A simulation activity (Gettysburg Address) was implemented in the lecture. The Gettysburg Address is an activity where students were given a speech by Nelson Mandela, were asked to circle any ten words, and then asked to find the average number of letters for each word in their sample. This was first done manually by recording, on paper, the words circled and the number of letters in each word, and then calculating the average number of letters. Students then shared their means and started to manually build a sampling distribution. They were then introduced to a simulation app, which selected ten words from the Gettysburg at random (without showing the actual words), and then calculated the mean automatically. They also produced 20-word samples, determined the means, and produced a distribution of the means. At the end of the lecture, students were provided with an online link to the extension tasks, which were the stimulus for the data collection in this study, and which are described in the results.

Data Analysis

The Structure of the Observed Learning Outcome (SOLO) taxonomy (Biggs & Collis, 2014) was used to analyse the data. SOLO is a qualitative model of assessment, established from existing cognitive models, which seeks to assess students' conceptual knowledge at varying levels of understanding. SOLO helps to separate the respondents from their responses, measuring only their knowledge or level of performance demonstrated at a particular time. This form of assessment has been used successfully to evaluate both summative and formative assessment in statistics education (e.g., Groth & Bergner, 2006; Watson et al., 2022). The SOLO taxonomy has five levels (Biggs & Collis, 2014), namely:

Prestructural. Responses at this level could be "I don't know", or involve merely reiterating the question, or have no relation to the question.

Unistructural. Responses might use a single cue from the question or relevant domain. There is no interrelationship of ideas.

Multistructural. Responses provided or interpreted at this level incorporate two or more aspects relevant to the questions, but they may not be interrelated, and conclusions in the response might be inconsistent.

Relational. Responses at this level weave together all connecting aspects in the response and make a coherent whole response to a given question.

Extended abstract. Responses at this level involve generalisation, set out principles on which responses are based, and may indicate application to other situations.

The students' responses were initially categorised against the SOLO levels by the first author. They were then categorised independently by the third and fourth authors, and any discrepancies were discussed and resolved.

Results

Comprehensive understanding of sampling distribution is evidenced by understanding of the sample mean, effect of sample size, variance in sample means, standard deviation (Watkins et al., 2014), standard error of the mean, and correct use of probability language (Chance et al., 2004). Presented below is each scenario with the corresponding questions, followed by a model response (sourced from https://askgoodquestions.blog/), a brief discussion about cogent words or description necessary for answering the question, general description of varying level of participants' responses using SOLO (Biggs & Collis, 2014), two examples of a given response, and a justification for the level of categorisation determined.

Scenario 1: Cats

> Assume that domestic housecats' body lengths (excluding the tail) have a mean of 18 cm and a standard variation of 3 cm.

Question 1. (Cat tail length longer than 20cm).

> Which is more likely—that a randomly chosen cat's length is longer than 20 cm, or that the average length of a randomly selected sample of 50 cats is more than 20 cm, or are these probabilities equal? Describe your thinking.

Model response. Since a length of 20 cm is longer than the mean and average values vary less than individual values, the likelihood of a cat exceeding 20 cm is higher than the likelihood of the sample average exceeding 20 cm.

Discussion. The model response implies variation (variation in the sample means, but greater variation in the individual values); when the mean of the whole population is 18 cm, the sample averages will be around this value, and thus likely to be less than 20 cm, whereas a single cat's length is more likely to be longer than 20 cm (correct use of probability language).

Out of eight respondents, most responses used the correct probability language, and most understood variation in sample mean. Three responses were classified as multistructural because they referred to multiple elements, such as the sample average, sample size, variation in mean, standard deviation and the correct use of probability language. One response was considered relational because it used mathematical calculations to back up its reasoning, and the other four were either pre-structural (e.g., gave vague responses, or just reiterated the question), or unistructural (only referred to the variation in mean or standard variation). Of particular interest here are two responses that were difficult to categorise.

> Response A: I think the probability of a sample of fifty cats having its mean fall at 20 cm is more likely than a single cat measuring exactly 20 cm, because 20 cm is only 1 standard deviation away.

Response A could be classified as a unistructural response or a pre-structural response. On one hand, it could be argued that the response was unistructural because it states a statistical claim with reference to one element of the problem (standard deviation). The claim, "because 20 cm is only 1 standard deviation away," shows that the respondent had some understanding of standard deviation, although the concept of standard deviation was not well-used. On the other hand, it could be said the response was pre-structural because the standard deviation is not relevant in this case. This

response was ultimately categorised as unistructural because the standard deviation was applied in the justification.

> Response B: The first probability is larger. The main difference is that when working with a sample, we use the standard error of the sample, which is equal to standard error divided by the square root of the sample size. p1 = p (> 20) = 1-NORM.DIST(20, 18, 3, 1) = 0.25; p2 = p(average length > 20; n = 50) = 1-NORM.DIST(20, 18, 3/sqrt 50, 1) = 0.0000012.

Response B was debated among the authors, with a multistructural categorisation initially considered, based on insufficient written explanation to address the instruction that said, "Describe your thinking." However, it was ultimately categorised as relational because the respondent justified the choice of standard error, chose the correct distribution, and then performed the correct calculations to justify the claim that the first probability was larger.

Question 2. (Cat tail length between 17–19 cm).

> Which is more likely—that a randomly chosen cat's length falls between 17 and 19 cm, or that an average length of 50 randomly chosen cats falls which this range, or are these probabilities equal? Describe your thinking.

Model response. Because the range is centered on the population mean, the probability of length being between 17 and 19cm is greater for a sample average than for an individual cat.

Discussion. In the analysis of this question, some responses related the sample mean to the population mean, which showed understanding of variation, sample mean, effect of sample size, and the use of correct probability language. Using the SOLO categorisations, most responses were at a unistructural level, and while two respondents realised that question one and question two were similar, the way this connection between question was stated brought about a question of where their responses should be placed on the SOLO categorisation scale.

> Response A: The same reason.

No other explanation was given, so if we go by the definition of the prestructural SOLO level, then, "The same reason" was considered prestructural. However, it is asking to take the justification for the other response (response given in question one) as the response for question two. If we go by the given response in question one "The second one should be larger generally as a more stable and closing to normal distributing data," which was categorised as unistructural, then we might categorise the response to question two as also being unistructural.

> Response B: Using the same logic as in the last question, the probability that the sample of 50 is between 17 and 19 cm will be higher.

Response B was categorised as multistructural, because the respondent was able to relate the first question to the second question using the phrase "using the same logic in the last question." In the previous question the participant wrote:

> The probability that the length of a randomly selected cat is no longer than 20cm is higher. The standard error in the sample of 50 will be lower than the standard deviation for one, as there is a 1/sqrt (50) multiplier. Thus, the confidence intervals will be smaller for the sample.

This response was also categorised as multistructural with respect to Question 1.

Scenario 2: Hospital Problem

Suppose that a region has two hospitals. Hospital A has about 10 births per day, and Hospital B has about 50 births per day. About 50% of all babies are boys, but the percentage who are boys varies at each hospital from day to day.

Question 3. Hospital problem.

> Over the course of a year, which hospital will have more days on which 60% or more of the births are boys—A, B, or negligible difference between A and B? Explain your option.

Model response. With a large sample, we would expect the mean proportion of boy births to be around 50%, whereas smaller samples will vary from this mean more often. Hospital A, which has fewer births (i.e., smaller sample size) will have more days with 60% or more of the births being boys because the smaller hospital will have more variation in the percentages of boys born on a day.

Discussion. The third question is a well-known statistical task in the field of introductory statistics. Out of eight respondents, most responses used the correct probability language. Three were prestructural because the question was only reworded or the response seemed like memorised theory with no depth to them, and the three unistructural responses only referred to variation in sample mean with no connection to other elements. Two individuals provided multi-structural responses to this question by explaining effect of sample size (see summary of responses in Table 1). Two responses of particular importance are recounted below.

> Response A: Hospital A. With less data, a smaller denominator of the standard error, means a larger range.

Response A could be determined to be at a prestructural level or a unistructural level of learning outcome. This is because only one useful piece of information was mentioned (standard error) which shows some indication of reasoning (unistructural) but this singular use of a term without adequate expression might arise as merely a memorised term (prestructural).

> Response B: Hospital A will have more variable days because of the low sample size.

Response B was categorised as multistructural as the respondent showed an understanding of variation and sample size, although the given response was not well expressed to show the implied relation between sample size of 10 versus 50. More connectivity between the two may have taken this response to the relational level. Also, the respondent could have been reiterating memorised theory, but it was not possible to confirm that possibility with the data available.

A summary of the responses for the three questions are presented in Table 1. Across the three questions and 24 responses in total, there were five prestructural, nine unistructural, nine multistructural, and one relational level responses.

Table 1

Number of Responses for Each SOLO Level

	SOLO Level			
	Prestructural	Unistructural	Multistructural	Relational
Question 1	1	3	3	1
Question 2	1	3	4	0
Question 3	3	3	2	0

Discussion

The analysis of the student responses demonstrated that participants generally understood the correct probability language that applied to the questions posed. Probability is one of the few topics in statistics studied from Years Foundation–12 in Australia (Australian Curriculum, Assessment and Reporting Authority, 2018), so it might be expected the participants have brought this familiarisation with the concept of probability from high school. This suggests that probability as an important idea in sampling distribution (Pfannkuch et al., 2012) could be an effective starting point for the teaching and learning of sampling distribution.

The five responses classified as either unistructural or multistructural support the observation that component elements of sampling distribution like "sample," "sampling," "variation," and "averages" (Watson, 2004; Watson et al., 2022) are vital to the understanding of bigger concepts. Lack of such understanding, however, may underpin the remaining prestructural responses, suggesting the misconception of variability between large and small samples (Ben-Zvi, 2004). Although not confirmed in the data presented here, we further postulate that higher levels of understandings could stem from using simulation as a pedagogical approach in teaching this concept (Chance et al., 2022).

Students' grasp of the ideas related to sampling distribution varied greatly, and using SOLO taxonomy proved helpful in distinguishing different levels of understanding. Higher order thinking abilities like analysis, synthesis, and assessment were encouraged by the form of the questions answered. However, only four of the SOLO taxonomy's five levels of learning outcomes were evidenced in the responses given. None of the questions achieved the abstract extended level of learning outcome, which may be because the questions were not expressed specifically to yield answers at that level.

Categorisation of the data using SOLO was at times problematic because responses were brief, non-existent, or ambiguous. This illustrates the problem of presentation or internalisation (Reaburn, 2014), as observed in the ways respondents wrote their responses. Since understanding of the concept can only be inferred based on what is written, lower-level responses could be due to a misunderstanding of the question wording, its presentation, or accompanying instructions. This suggests additional data, such as from an interview, is needed to support the SOLO classifications. Interviews using open-ended questions would provide participants the opportunity to demonstrate understanding at the extended abstract level described by Groth and Bergner (2006). The results in Table 1 shows the majority of responses were unistructural or multistructural, with only one relational response This suggests that the simulation activity was not fully successful in supporting students' conceptual understanding for the extension tasks, certainly in making the connections between the simulation observations and different contexts.

Conclusion

The study reported in this paper differs from previous research in that it has looked at the understanding of the fundamental components of sampling distribution, rather than the more common focus on sampling distribution for statistical inference. The evidence presented here serves to shed light on the elements of sampling distribution best understood by students, and the levels of understanding observed in their responses. It highlights the complexity of sampling distribution and suggests that achieving sophisticated understanding of the concept is possible if the fundamental components of the concept are well-understood. The results from this report could potentially be used to inform a methodological approach in analysing students' understanding of fundamental components of a statistical concept.

References

Australian Curriculum, Assessment and Reporting Authority. (2018). *The Australian curriculum: Mathematics* (senior secondary). ACARA. https://www.australiancurriculum.edu.au/senior-secondary-curriculum/mathematics/

Ben-Zvi, D. (2004). Reasoning about variability in comparing distributions. *Statistics Education Research Journal*, *3*(2), 42–63. https://doi.org/10.52041/serj.v3i2.547

Biggs, J. B., & Collis, K. F. (2014). *Evaluating the quality of learning: The SOLO taxonomy (structure of the observed learning outcome)*. Academic Press.

Blejec, A. (2003). Teaching statistics by using simulations on the Internet. In *Statistics and the Internet. Proceedings of the IASE/ISI satellite conference*, Berlin, Germany. https://iase-web.org/documents/papers/sat2003/Blejec.pdf?1402524992

Bruce, P. C. (2014). *Introductory statistics and analytics: A resampling perspective*. John Wiley & Sons.

Chance, B., delMas, R., & Garfield, J. (2004). Reasoning about sampling distribitions. In D. Ben-Zvi, & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 295–323). Springer. https://doi.org/https://doi.org/10.1007/1-4020-2278-6_13

Chance, B., Tintle, N., Reynolds, S., Patel, A., Chan, K., & Leader, S. (2022). Student performance in curricula centred on simulation-based inference. *Statistics Education Research Journal, 21*(3), Article 4. https://doi.org/10.52041/serj.v21i3.6

Duckworth, W. M., & Stephenson, W. R. (2002). Beyond traditional statistical methods. *The American Statistician, 56*(3), 230–233. https://doi.org/10.1198/000313002173

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

Groth, R. E., & Bergner, J. A. (2006). Preservice elementary teachers' conceptual and procedural knowledge of mean, median, and mode. *Mathematical Thinking and Learning, 8*(1), 37–63. https://doi.org/10.1207/s15327833mtl0801_3

Hesterberg, T. (2015). What teachers should know about the bootstrap: Resampling in the undergraduate statistics curriculum. *The American Statistician, 69*(4) 371–386, http://doi.org/10.1080/00031305.2015.1089789

Horton, N. J. (2015). Challenges and opportunities for statistics and statistical education: Looking back, looking forward. *The American Statistician, 69*(2), 138–145. https://doi.org/10.1080/00031305.2015.1032435

Kula, F., & Koçer, R. G. (2020). Why is it difficult to understand statistical inference? Reflections on the opposing directions of construction and application of inference framework. *Teaching Mathematics and its Applications: An International Journal of the IMA, 39*(4), 248–265. https://doi.org/10.1093/teamat/hrz014

Larwin, K. H., & Larwin, D. A. (2011). Evaluating the use of random distribution theory to introduce statistical inference concepts to business students. *Journal of Education for Business, 86*(1), 1–9. https://doi.org/10.1080/08832321003604920

Mackenzie, N., & Knipe, S. (2006). Research dilemmas: Paradigms, methods and methodology. *Issues in Educational Research, 16*, 193–205. https://www.iier.org.au/iier16/mackenzie.html

Makar, K., & Rubin, A. (2018). Learning about statistical inference. In D. BenZvi, J. Garfield, & K. Makar (Eds.), *International handbook of research in statistics education* (pp. 261–294). Springer International Publishing.

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine, 38*(11), 2074–2102. https://doi.org/10.1002/sim.8086

Ozmen, Z. M., & Guven, B. (2019). Evaluating students' conceptual and procedural understanding of sampling distributions. *International Journal of Mathematical Education in Science and Technology, 50*(1), 25–45.

Pfannkuch, M., Wild, C. J., & Parsonage, R. (2012). A conceptual pathway to confidence intervals. *ZDM-Mathematics Education, 44*(7), 899–911. https://doi.org/10.1007/s11858-012-0446-6

Reaburn, R. (2014). Introductory statistics course tertiary students' understanding of *p*-values. *Statistics Education Research Journal, 13*(1), 53–65. https://doi.org/10.52041/serj.v13i1.298

Rossman, A. J., & Chance, B. L. (2014). Using simulation-based inference for learning introductory statistics. *Wiley Interdisciplinary Reviews: Computational Statistics, 6*(4), 211–221. https://doi.org/10.1002/wics.1302

Sigal, M. J., & Chalmers, R. P. (2016). Play it again: Teaching statistics with Monte Carlo simulation. *Journal of Statistics Education, 24(*3), 136–156. https://doi.org/10.1080/10691898.2016.1246953

Sued, M., & Valdora, M. (2021). Each student with her/his own data: Understanding sampling distributions. *arXiv,* Article 2105.02727. https://doi.org/10.48550/arxiv.2105.02727

Watkins, A. E., Bargagliotti, A., & Franklin, C. (2014). Simulation of the sampling distribution of the mean can mislead. *Journal of Statistics Education, 22*(3). https://doi.org/10.1080/10691898.2014.11889716

Watson, J. M. (2004). Developing reasoning about samples. In D. Ben-Zvi, & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 277–294). Springer.

Watson, J., Wright, S., Fitzallen, N., & Kelly, B. (2022). Consolidating understanding of variation as part of STEM: Experimenting with plant growth. *Mathematics Education Research Journal*. https://doi.org/10.1007/s13394-022-00421-1

Watts, D. G. (1991). Why is introductory statistics difficult to learn? And what can we do to make it easier? The *American Statistician, 45*(4), 290–291. https://doi.org/10.2307/2684456

Zieffler, A., Garfield, J., & Fry, E. (2018). What is statistics education? In D. BenZvi, J. Garfield, & K. Makar (Eds.), *International handbook of research in statistics education* (pp. 37–70). Springer International Publishing. https://doi.org/10.1007/978-3-319-66