International Journal of Learning, Teaching and Educational Research
Vol. 4, No.1, pp. 51-60, April 2014

Investigating Reliability and Validity for the Construct of Inferential Statistics

Saras Krishnan and Noraini Idris
University of Malaya
Kuala Lumpur, Malaysia

Abstract.A hierarchical construct to assess and describe students' learning of inferential statistics has been previously developed using the Rasch analysis. In particular, the Rasch Partial Credit Model was instrumental in identifying the number of strata in the construct and in establishing the reliability and validity of the instrument used. In this paper, the analysis is replicated with a different sample of students to investigate if the reliability and validity still hold.

Keywords: assessment; Partial Credit Model; Rasch analysis

Introduction

Past studies in various aspects of inferential statistics provide evidence of students' continual difficulties in learning the many aspects and concepts of inferential statistics (e.g., Francis, Kokonis,& Lipson, 2007; Weinberg, Wiesner,& Pfaff, 2010). This situation is worrying since inferential statistics is taught in a majority of courses,and the knowledge and skills of inferential statistics will be required at one time or another by the students. Despite the many studies of students' learning of various topics of inferential statistics, at present there is need for more research in this area(Smith, 2008; Sotos, Vanhoof, Noortgate,& Onghena, 2009).

A construct of learning to describe students' understanding of inferential statistics in hierarchical levels has been developed as part of the main author's postgraduate research. The developmental process of this construct is discussed in Krishnan and Idris (2013a). Discussion included the use of Rasch analysis in establishing the reliability and the validity of the results, and in determining the number of levels in the construct. Further, another paper discussed the use of thehierarchical construct to investigate students' learning of inferential statistics (Krishnan & Idris, 2013b). In this paper, we investigate the reliability and validity of the results using a different sample of students with the same sample size. The purpose of this investigation is to determine if the conditions of reliability and validity are still fulfilled when a different sample of students is used.

Literature Review

Issues of concernin the assessments in statistics education included the assessment of various statistics topics and the assessment of aspects of statistics that are indicative of students' varying levels of understanding (Bude, 2006). Another concern is that assessment of students' statistical learning is yet to be adequately addressed (Smith, 2008). With regards to these concerns, several attempts have been made to assess and describe students' learning of statistics in hierarchical stages of understanding.

The statistical literacy construct developed by Watson and Callingham (2003) describes students' understanding of statistics involving average and chance, sampling and inference, representation of data, and variation. The two frameworks for the statistical literacy construct are Biggs and Collis' (1982, 1991) SOLO Taxonomy and Watson's (1997) three tier statistical literacy model. Other constructs that assessed students' learning of statistics have basically evolved from Watson and Callingham's construct (e.g., Callingham, 2009; Kaplan & Thorpe, 2010; Watson, Kelly,& Izard, 2005). On the other hand, Kataoka, da Silva, Vendramini, and Cazorla (n.d.) used the SOLO Taxonomy to categorize students' responses to a statistics questionnaire but did not offer any learning construct in their study.

The Construct of Inferential Statistics (Krishnan & Idris, 2013b)contains six hierarchical levels that describe students' understanding of inferential statistics in increasing complexity as we ascend the levels. For instance, the first level involves students' ability to identify inferential terminologies and symbols when presented in contextual form while the fifth level involves understanding of sampling, students' ability to infer in different contexts, and knowledge and understanding of inferential procedures and concepts.

The different constructs of statistics have been developed using the Rasch model for analysis of data. Rasch model has been particularly useful in statistics assessments in determining students' levels of understanding of various statistics concepts.Apart from that, Rasch analysis has also been used to investigate students' understanding of basic statistical concepts (Kassim, Ismail, Mahmud,& Zainol, 2010),and to investigate attitude and knowledge of statistics among postgraduate students (Mahmud, 2011).

There are two important reasons for using the Rasch analysis in our studies. First, Rasch analysis is used to determine the number of strata or levels of the construct in describing the stages of students' understanding of inferential statistics. Second, Rasch analysis is used to establish the reliability and validity of the instrument and the sample of students. The first reason is facilitated by the use of the item separation reliability and the item-person map. The second reason is facilitated by the use of the fit analysis primarily the table of summary statistics and the table of misfit order of items. Explanation on these can be found in Krishnan and Idris (2013a, 2013b). The item-person maps are not included in this paper due to the irrelevancy to the discussion here.

Among the different types of Rasch models, the Partial Credit Model (Masters, 1982) is especially instrumental in our research because it accommodates items that have different hierarchical scoring categories. In other words, the Rasch Partial Credit Model allows the dichotomous and polytomous items to be put together in the same instrument (Bond & Fox, 2007). Thus, it is a model particularly practical and instrumental in education assessments because it is common for students to provide partly correct answers to any questions in a written assessment.

Methodology

Research design

Descriptive research design, in particular the cross-sectional survey method was used to collect quantitative data. Descriptive research primarily describes a current state of affairs usually with the use of visual aids (Knupfer & McLellan, 2001). Our studies employed the descriptive research design because we want to describe categories of information relating to students' understanding of inferential statistics with the aid of the item-person map in the Rasch analysis.

Instrumentation

The instrument used to collect data in this study is a task-based questionnaire on inferential statistics. Three progressive sets of pilot studies were conducted in developing the instrument. At each stage the instrument was further improved to meet the criteria of Rasch analysis particularly in terms of the reliability and validity of the instrument. In addition, the language and the structure of the questions were also modified to be able to elicit more valid responses from the students.

The purpose of the first pilot study was to collect baseline data to get an idea of the possible responses to the questionnaire and possible problems in coding these responses. The second and third pilot studies had a more definite purpose of investigating the quality of the instrument whereby items that do not meet the conditions of reliability and validity are either removed from the instrument or are restructured. The results of these pilot studies are reported in Krishnan and Idris (2013a).

The final instrument named as the Questionnaire for the Construct of Inferential Statistics contained 10 main items and 21 items altogether and is a task-based questionnaire that allows students to give open-ended responses. As such, we are able to gather a multitude of different responses and can perceive a greater variability of students' learning of inferential statistics in the higher education. As of now, we are not able to furnish the questionnaire due to the unpublished status of the first author's thesis.

Data collection

The actual data collection process was carried out over a period of 6 months.Two factors contributing to the duration is the availability of the students and authorization from the higher education institutions in concern.

Each data collection required 40 minutes where in the first 10 minutes the students were briefed about the purpose of the data collection and were given the necessary instructions. Then, students had 30 minutes to respond to the items in the questionnaire individually. Data collection involved 150 students in each sample. In using Rasch analysis, there are no specific requirements for the sample size. In general, the sample size is large enough if the item reliability is not less than 0.90.

Samples of study

Malaysia is a country in the South East Asia with a population of various ethnic, cultural and lingual backgrounds. The many ethnic groups predominantly consist of the Malay, Chinese and Indian races. The national language is the Malay language while English is widely used as the second language. The two main higher education providers in Malaysia are the government (60%) and the private sector (40%). Notwithstanding, the number of students opting for a private education has been increasing over the years (Krishnan & Idris, 2013c).

Purposive sampling has been used to identify the samples of students from the different higher education institutions. Sample 1 is made up of students from one private and one semi-private higher education institutions from two different states in the central region of the country. The private higher education institution was founded more than a quarter century ago and at present offers a range of programs from pre-university studies to postgraduate courses. The students for this study are taken from one pre-university program and two different degree programs from this private higher education institution.

The semi-private higher education institution has been in operation longer than the private higher education institution, having evolved from a training centre to a full fledge higher education provider. Some of the courses available at this institution are architecture, communication studies and dentistry. The students for this study are taken from an external pre-university program at this semi-private higher education institution, which is a different pre-university program than the one from the private higher education institution.

Sample 2 consists of students from a public higher education institution in the northern region that offers a range of undergraduate and postgraduate programs in pedagogy in the different faculties it houses. The 150 students sampled from this higher education institution belong to the same diploma program and is taught by the same instructor in four separate classes. Although the official medium of instruction at this institution is English, the Malay language was often used because the students are largely from the Malay language speaking background and thus have limited English speaking and writing capabilities. The teaching materials too are sometimes provided in dual languages to compensate students' English language inadequacy.

The defining differences between these two samples are: (i) gender, (ii) ethnicity, and (iii) English language capability. Table 1 shows the composition of students in the samples according to this segregation. In comparison, both samples have

more female students than the male students. The largest ethnic group for Sample 1 is Chinese while the largest ethnic group for Sample 2 is Malay. On the other hand, the smallest ethnic group for Sample 1 is other ethnicity while the smallest ethnic group for Sample 2 is Indian. Further, a small percentage of the students in Sample 1 maintained that they have good English speaking and writing capabilities whereas for Sample 2 the students' English capability ranged from moderate to poor. None of the students in Sample 2 have good English speaking or writing capability. In fact, for both samples, the largest percentages of students have moderate speaking and writing capabilities of the English language.

Table 1: Composition of students in the samples

		Sample 1	Sample 2
Gender	Male	42.7%	25.3%
	Female	57.3%	74.7%
Ethnicity	Malay	29.3%	88.7%
	Chinese	56%	2%
	Indian	8%	0.7%
	Others	6.7%	8.7%
Spoken English	Good	14.7%	0%
	Moderate	67.3%	81.3%
	Poor	18%	18.7%
Written English	Good	15.3%	0%
	Moderate	66%	87.3%
	Poor	18.7%	12.7%

Analysis of Results

Table 2 shows the reliability and fit indices for Sample 1. These results have been discussed in earlier paper that described the development of the hierarchical construct (Krishnan & Idris, 2013b). The purpose of this study is to investigate the results of these indices for a different sample, Sample 2. The item separation reliability determines the breadth of the items whereby a value more than 1.00 indicates that the items have enough breadth as with the case of Sample 1. In a similar manner, the person separation reliability must be more than 1.00 to warrant that the students are measured across the continuum. This condition has been met by Sample 1.

Table 2: Reliability and fit indices for Sample 1

Item separation reliability	6.48
Item infit mean square	1.00 (s.d. 0.08)
Item reliability	0.98
Person separation reliability	1.77
Person infit mean square	1.03 (s.d. 0.33)
Person reliability	0.76
Cronbach's alpha	0.75

The item infit mean square and the person infit mean square must be in the range of 1.00 to 1.20 to be reckoned as acceptable. Value less than 1.00 means that the responses are too predictable. It also suggests the presence of redundant items. On the other hand, value more than 1.20 suggests unpredictable responses or inappropriate response patterns. Meanwhile, the standard deviation must be smaller than 2.00 to indicate little misfit.Both the item infit mean square and the person infit mean square for Sample 1 as well as their standard deviation values met the required conditions.

As mentioned in Krishnan and Idris (2013b) there is no hard and fast rule on the acceptable range of the fit statistics and different researchers have complied with different ranges of these values. Discussion on the possible different values of the fit statistics can be found in Green and Frantom (2002), and Linacre (2002).
In addition, the item reliability, the person reliability and Cronbach's alpha in Table 1 are more than 0.70. The item reliability and the person reliability values are equivalent to the value of Cronbach's alpha, said Green and Frantom (2002). In this study, Cronbach's alpha of 0.70 is used as an acceptable reliability coefficient (Nunnaly, 1978; Santos, 1999).

For Sample 2, some of the aforementioned conditions were met whereas others were not. First, the item separation reliability of 3.98 and the person separation reliability of 1.03 both satisfy the condition that these values must be more than 1.00. However, they are lower than the values for Sample 1. This observation suggests that the spread of the items and students in Sample 2 is smaller compared to Sample 1. The item infit mean square for Sample 2 is in the stipulated range of between 1.00 and 1.20 but the person infit mean square does not fulfil this condition. Likewise, the item reliability is more than 0.70 but the person reliability is not. The Cronbach's alpha too does not meet the condition of reliability.

Table 3: Reliability and fit indices for Sample 2

Item separation reliability	3.98
Item infit mean square	1.00 (s.d. 0.10)
Item reliability	0.94
Person separation reliability	1.03
Person infit mean square	0.98 (s.d. 0.45)
Person reliability	0.51
Cronbach's alpha	0.58

The misfit order of items for analysis of both samples is displayed in Table 4. The infit mean square values (denoted by MNSQ) and the infit z-standardized values (denoted by ZSTD) are investigated to establish the validity of an instrument whereby the conditions for validity are:
(i) MNSQvalues between 0.70 and 1.33 (Watson & Callingham, 2003), and
(ii) ZSTD values between -2.00 and +2.00 for samples of sizes between n = 30 and n = 300 (Bond & Fox, 2007).

Table 4 shows that both conditions have been met for Sample 1 and Sample 2. For Sample 1 the MNSQ values ranged from 0.86 to 1.17 and the ZSTD values ranged from -1.60 to 1.40 (Krishnan & Idris, 2103b). Meanwhile, for Sample 2 the MNSQ values ranged from 0.79 to 1.27 and the ZSTD valuesranged from -0.80 to 1.40.

Table 4: Misfit order of items

Sample 1		*Sample 2*	
MNSQ	*ZSTD*	*MNSQ*	*ZSTD*
1.15	1.40	0.92	-0.20
1.12	1.30	1.27	1.00
1.17	1.20	1.07	0.90
1.07	0.60	1.14	1.40
1.13	1.10	1.09	0.60
0.94	-0.60	1.00	0.10
0.94	-0.60	1.07	0.30
1.00	0.00	1.04	0.70
0.92	-0.40	1.04	0.40
1.03	0.40	1.01	0.20
1.01	0.20	1.02	0.20
0.96	-0.40	1.01	0.10
0.99	-0.10	1.00	0.30
0.99	-0.10	1.00	0.30
0.98	-0.10	0.99	-0.10
0.98	-0.10	0.92	-0.80
0.97	-0.20	0.98	0.00
0.94	-1.20	0.96	-0.10
0.91	-0.90	0.89	-1.30
0.88	-1.10	0.89	-0.40
0.86	-1.60	0.79	-0.60

Overall, the analyses from Sample 1 and Sample 2 reveal that the reliability and validity of the instrument has been established regardless of the sample diversity. Especially the results of analysis of Sample 2 corroborate the quality of the Questionnaire for the Construct of Inferential Statistics because the conditions of reliability and validity have been met by the instrument despite the students in Sample 2 not fulfilling the conditions of reliability.

Conclusion

Statistics assessment has evolved in the past 40 years (Jolliffe, 2007) from assessing students' knowledge of statistical formulas to assessing students' understanding of statistical concepts. The various existing constructs to assess students' learning of statistics are largely concerned with students' understanding of the descriptive statistics. We have developed a construct to assess students' learning of the inferential statistics in the higher education contexts and have discussed the development of this construct in earlier papers (Krishnan & Idris, 2013a, 2013b).

The development of the construct of inferential statistics basically supports the requirement to increase the number of literature in the area of students' learning and understanding of inferential statistics because studies in this area are still scarce at present (Smith, 2008).The construct of inferential statistics can be utilized by statistics educators to improve students' understanding of the logic of statistical investigations and the need to infer from samples to populations. It can also aid in developing students' deep and connected understanding of inferential statistics. By identifying the different levels of students' understanding of inferential statistics, instructors can focus on the development and improvement of students' understanding of the levels in concern.

In this paper, we investigated if the reliability and validity achieved earlier is maintained if a different sample of students were used. It was found that although the sample of students in this study did not meet some conditions of reliability, the reliability and validity of the instrument was fulfilled. This served to verify the validation of the instrument and subsequently the validation of the construct. However, this study did not investigate if gender, ethnicity or language capability or a combination of them affected the results. We suspect language or rather lack of it could have played a major role because a number of questionnaires from Sample 2 were barely attempted and some students conceded that it was because they did not know how to explain their reasoning in English.

Lesser (2010) believes that student diversity interacts with the learning of statistics and it is important for instructors to use student diversity as an opportunity instead of obstacle. Lesser and Winsor (2009) also believe that language is an important factor in students' performance but found that at present there is lack of research on statistics learning involving English articulateness. Future possible work with respect to this paper is to investigate in detail how students' different language capabilities affect the reliability and validity of the construct.

References

Biggs, J.B.,& Collis, K.F. (1982). *Evaluating the quality of learning: The SOLO taxonomy*. New York, NY: Academic Press.

Biggs, J.B.,& Collis, K.F. (1991). Multimodal learning and the quality of intelligent behaviour. *Intelligence: Reconceptualization and measurement*, 57-76.

Bond, T.G.,& Fox, C.M. (2007). *Applying The Rasch Model: Fundamental Measurement in the Human Science*. Mahwah, NJ: Lawrence Erlbaum Associates.

Bude, L. (2006). Assessing students' understanding of statistics. In *Proceedings of the 7th Annual Meeting of the International Conference on Teaching Statistics.* Brazil: ISI. Retrieved April 16, 2010 from www.stat.auckland.ac.nz/~iase/publications/17/6G3_BUDE.pdf

Callingham, R. (2009). Using Rasch Measurement to Identify Cross-cultural Aspects of Statistical Literacy.*Changing Climates: Education for Sustainable Futures*, *1*, 1-10.

Francis, G., Kokonis, S.,& Lipson, K. (2007). Enhancing Student Understanding in Statistical Inference-Assessing the Effectiveness of a Computer Interaction. *IASE/ISI Satellite*. Retrieved February 20, 2011 from http://iase-web.org/documents/papers/sat2007/Francis_et_al.pdf

Green, K. E., & Frantom, C. G. (2002). Survey Development and validation with the Rasch Model. Paper presented at the*International Conference on Questionnaire Development, Evaluation, and Testing*. Charleston, SC.Retrieved February 6, 2013 from http://www.jpsm.umd.edu/gdet/final_pdf_papers/green.pdf

Jolliffe, F. (2007). The Changing Brave New World Of Statistics Assessment. *IASE/ISI Satellite*. Retrieved February 20, 2011 from http://www.stat.auckland.ac.nz/~iase/publications/sat07/Jolliffe.pdf

Kaplan, J.J.,& Thorpe, J. (2010). Post Secondary and Adult Statistical Literacy: Assessing Beyond the Classroom. In *Proceedings of the Eighth International Conference on Teaching Statistics.* Netherlands: ISI.Retrieved September 2, 2011 from http://iase-web.org/documents/papers/icots8/ICOTS8_5E3_KAPLAN.pdf

Kassim, N.A., Ismail, N.Z., Mahmud, Z., & Zainol, M.S. (2010). Measuring Students Understanding of Statistical Concepts using Rasch Measurement. *International Journal of Innovation, Management and Technology*, *1*(1), 13-19.

Kataoka, V.Y., da Silva, C.B., Vendramini, C.,& Cazorla, I. (n.d.). *Using Rasch Partial Credit Model to analyze the Responses of Brazilian Undergraduate students to a Statistics Questionnaire*. Retrieved June 1, 2014 from http://www.cerme7.univ.rzeszow.pl

Knupfer, N.N.,& McLellan, H. (2001). Descriptive Research Methodologies. In D.H. Jonassen (Ed.), *Handbook of Research for Educational Communications and Technology* (pp. 1196-1213). Mahwah, NJ: Lawrence Erlbaum Associates.

Krishnan, S., & Idris, N. (2013a).The Development of an Assessment Construct for Inferential Statistics. Paper presented at the *International Conference on Assessment for Higher Education Across Domains and Skills (AHEADS2013)*. Kuala Lumpur, Malaysia.

Krishnan, S., & Idris, N. (2013b). The Use of a Hierarchical Construct to Investigate Students' Learning of Inferential Statistics. In *Proceedings of the Joint IASE/IAOS Satellite Conference*(pp. 1-8).Macao, China. August 2013.

Krishnan, S., & Idris, N. (2013c). The Use of Graphics Calculator in a Matriculation Statistics Classroom: A Malaysian Perspective. *Technology Innovations in Statistics Education*, *7*(2), 1-13.

Lesser, L.M. (2010). Equity and the Increasingly Diverse Tertiary Student Population: Challenges and Opportunities in Statistics Education. In *Proceedings of the Eighth International Conference on Teaching Statistics* (pp. 1-6).Netherlands: ISI. Retrieved December 10, 2013 from http://iase-web.org/documents/papers/icots8/ICOTS8_3G3_LESSER.pdf

Lesser, L.M., & Winsor, M.S. (2009). English language learners in introductory statistics: Lessons learned from an exploratory case study of two pre-service teachers. *Statistics Education Research Journal, 8*(2), 5-32.

Linacre, J.M. (2002). What do Infit and Outfit, Mean-Square and Standardized Mean. *Rasch Measurement Transactions*, *16*(2).

Mahmud, Z. (2011). Diagnosis of Perceived Attitude, Importance, and Knowledge in Statistics Based on Rasch Probabilistic Model. *International Journal of Applied Mathematics and Informatics*, *5*, 291-298.

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Pschometrika*, *47*, 149-174.

Nunnaly, J. (1978). *Psychometric theory*. New York, NY: McGraw Hill.

Santos, J.R.A. (1999). Cronbach's alpha: A tool for assessing the reliability of scales. *Journal of extension*, *37*(2), 1-5.

Smith, T.M. (2008). *An Investigation into Student Understanding of Statistical Hypothesis Testing*. Doctoral dissertation. University of Maryland.

Sotos, C., Vanhoof, S., Noortgate, W.,& Onghena, P. (2009). How confident are students in their Misconceptions about Hypothesis Tests?. *Journal of Statistics Education*, *17*(2).

Watson, J.M. (1997). Assessing statistical literacy using the media. In I. Gal & J.B. Garfield (Eds.),*The Assessment Challenge in Statistics Education* (pp.107-121). Netherlands: IOS Press.

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

Watson, J.M., Kelly, B.A.,& Izard, J.F. (2005). Statistical Literacy over a Decade. *Building connections: Theory, research and practice*, *1*, 775-782.

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.