Applying the Paek et al. method for calculating over- and under-confidence at the item and test levels

Aaron Olaf Batty
abatty@sfc.keio.ac.jp
Keio University
Lancaster University

Abstract

This article explains investigating over- and under-confidence on tests and test items using the method developed by Paek et al. (2008) for use with Rasch and other IRT measures. The data for this demonstration originated from a study of 199 Japanese high school and university students, investigating their knowledge of a number of special uses of verbs of utterance in English. The paper provides practical information on the calculations necessary for the use of the Paek et al. method under the Rasch model and the interpretation of the results. Finally, the same data are scaled with the two-parameter IRT model and the Paek et al. method is applied for comparison.

Keywords: confidence, accuracy, discrepancy, overconfidence, underconfidence, Rasch, two-parameter IRT

Many researchers have attempted to enhance their test data by incorporating a confidence scale to their instruments. Such data can prove useful within an instructional context, as discrepancy between confidence and accuracy signals a need for further instruction at a finer level of detail than simple accuracy (i.e., right/wrong) data can. On such instruments, the respondent typically responds to answers an item and then indicates his or her confidence of the accuracy of the answer on a separate scale. Such instruments are by no means new, and methods for interpreting their results have been in use for decades (e.g. Hakstian & Kansup, 1975; Keren, 1991; Lichtenstein, Fischhoff, & Phillips, 1981), but a method for incorporating confidence data with IRT measures was not developed until more recently.

Paek et al. (2008) describe a method of comparing the accuracy and the confidence of respondents on items and on tests overall, taking advantage of Rasch (1960) IRT measures to control for overall accuracy. It allows the researcher to examine the amount of over- or under-confidence on each item, and to compare it to overall ability on the measure. The method is described in detail in their 2008 ETS research report, but the discussion is not very accessible to most readers in the field of language testing. In this brief article, I will explain the process in greater detail, in the hopes that more language testing researchers can add it to their repertoire of test analyses.

Although a Rasch program is necessary to obtain the difficulty/ability estimates, the remainder of the calculations can be done in a spreadsheet application such as Microsoft Excel. A high-quality graphing software package is necessary to produce the plots discussed at the end of the article.

Data

The data used for illustration of this method are from the continuation of the utterance verb study described in Sato and Batty (2012). The participants were 199 Japanese learners of English, ranging from high school through undergraduate university, with abilities ranging from beginner through bilingual. The items were 20 gap-fill sentences, for which the respondent must choose an appropriate verb of utterance (*speak, talk, say,* or *tell*) and then report his/her confidence of the correctness of his/her answer on a scale from 1 (not confident) through 3 (very confident). Hence, every item has both a dichotomous accuracy variable as well as a polytomous confidence variable.

Method

The method compares values that Paek et al. refer to as the *objective probability* and the *subjective probability*. The former is simply the probability of a given respondent answering a given item correctly, given his/her level, while the latter is the probable confidence level of a person of that level, on that item. The former is conceptualized as a measure of *accuracy*, and the latter of *confidence.*

First, a metric of over/underconfidence must be obtained for each person. This is calculated as the difference between an individual's subjective probability and objective probability, where a positive value (indicating that the examinee's subjective probability is larger than his/her objective probability—or, to put it another way, his/her confidence outstripped his/her ability) is interpreted as overconfidence, and a negative as underconfidence, as below (Paek et al., 2008, p. 3):

$$\text{Overconfidence} = P_i^* - P_i > 0 \tag{1}$$

$$\text{Underconfidence} = P_i^* - P_i < 0 \tag{2}$$

where P_i is the objective probability of item i (whatever it may be in this case, e.g., the objective probability for a certain person of a certain ability on item 15 would be P_{15}) and P^*_i is the subjective probability of item i. The next section explains how to obtain these values.

Obtaining the objective probability scores (P_i)

The first step is to run Rasch estimation on the test. All Rasch packages provide an estimate of the probability of success for each person on each item, although most packages call this statistic *expected score.* It is critical to note, that this is not the probability of the respondent's actual response to the item; it is the probability of the respondent answering the item in question correctly, given his/her overall score and the difficulty of the item.

Following are brief instructions for obtaining the expected score statistic in two popular Rasch estimation packages.

Winsteps

In Winsteps (Linacre, 2012a), expected scores are found in the observation file (aka XFILE) available as an output from the "Output Files" menu. The output will have one line per person, per item. Therefore, for example, if your test has 10 items and you had 25 respondents, there would be 250 lines.

The probability of a correct response by the person on the item is found in the "EXPECTED" column. This is the objective probability for calculating over/underconfidence.

Facets

In Facets (Linacre, 2012b), expected scores are found in the "Residuals/Responses file", available from the "Output Files" menu. The expected scores are found in the column labeled "Exp".

Obtaining the subjective probability scores (P^*_i)

The subjective probability scores are simply the average confidence scores for all the examinees at a certain Rasch ability level, on each item, rescaled to have a range of 0 to 1 (i.e., the range of a probability, regardless of how long the confidence scale on the test in question was).

For example, in the example utterance verb data, the average confidence score on item 1, for all the respondents with a Rasch ability score (denoted θ in mathematical notation; here it is denoted θ_{ac} because it is the ability score on the *accuracy* dimension) of 0.52 is 1.62. This is divided by 3 to rescale it to a number from 0 to 1, representing a probability: 0.54. This rescaled value is the subjective probability for item 1 for every person with a Rasch ability score of 0.52. Because Rasch ability scores are scaled from total raw scores, the number of discrete Rasch scores in any dataset is finite, and, furthermore, is rather small. In the case of the example instrument with 20 items, there are only 21 possible Rasch scores (i.e., one for each number of items correct, from 0 through 20). In practice, however, there were only 18 discrete Rasch ability scores for the respondents.

There are many ways to quickly calculate the subjective probability, but the way I approached it was by using using Excel's VLOOKUP formula to match Rasch scores from the score file to respondents in a worksheet with their objective probabilities. I then used VLOOKUP on another column to bring in the confidence score for each item for each person. Finally, I used the AVERAGEIFS function to average the confidence scores for each item, for each Rasch score (θ_{ac}).

To calculate the subjective probability for the first respondent on the first item in Figure 1, for example, I used the following Excel formula:

```
=AVERAGEIFS(D:D,C:C,C2,B:B,B2)/3
```

which instructs Excel to average values from the D column if the value in the C column ("Ability", or the Rasch ability score for that respondent) matched the value for this person (C2), and if the value in the B column (the item number) matched the value for this particular row (B2). Finally, it divides that average by 3 to rescale it to 0 – 1. This results in a rescaled average of all the confidence scores for people with the same ability score as respondent W007 (i.e., 0.52) on the item in question (item 1).

	A	B	C	D	E	F
1	Examinee	Item#	Ability	Confidence	Obj Prob	Sbj Prob
2	W007	1	0.52	2	0.605	0.540
3	W007	2	0.52	2	0.340	0.575
4	W007	3	0.52	2	0.260	0.632
5	W007	4	0.52	2	0.775	0.621
6	W007	5	0.52	3	0.845	0.713

Figure 1. Calculating the subjective probability.

Calculating the over/underconfidence scores

To calculate the over/underconfidence score for each person and each item, simply subtract the objective probability (aka the Rasch expected score) from the subjective probabilities. Note, however, that the resulting over/underconfidence score will be the same for all people who share the same Rasch ability score (θ_{ac}), on each item. From this point on, the Paek et al. method does not concern itself with individual examinees. For this reason, it is advisable to paste the data into a new sheet and remove the duplicates so that what remains is one list of all of the numbers for each Rasch ability estimate level.

Calculating the Item Discrepancy Index (*IDI*)

The IDI is a summary statistic for the amount and direction of the discrepancy between accuracy (the objective probability) and confidence (the subjective probability), also called the over/underconfidence, on each item of the instrument. It is weighted by the ratio of examinees at each Rasch ability level. This controls for the distribution of ability levels by making more-common ability levels contribute more than less-common ones. It is calculated as below (Paek et al., 2008, p. 4):

$$(IDI_i) = \sum_{\theta_{ac}} (P_i^* - P_i) w(\theta_{ac}) \tag{3}$$

where:

$$w(\theta_{ac}) = \frac{N_{\theta_{ac}}}{N} \tag{4}$$

Calculating the weighting statistics

The first half of equation 3 has already been calculated at this point, as it is simply the over/underconfidence statistic described above. The next calculation is the weighting variable described in equation 4, which is calculated by doing nothing more that counting the number of people at a certain Rasch ability score level (θ_{ac}), and dividing it by the total N of the sample. In a spreadsheet program, this can be accomplished by using the COUNTIF and COUNT functions. The example in Figure 2 uses the following formula to calculate the weighting statistic in the cell F2:

```
=COUNTIF(Data!C:C,'IDI Calcs'!B2)/COUNT(Data!C:C)
```

It first counts up all the rows in the sheet with the full dataset that have the same ability score (θ_{ac}) as appears in B2. It then divides that count by the count of all the rows in the full dataset. Because each person has exactly 20 rows of data, it does not matter that we are counting items, rather than people. The result will be the same.

	A	B	C	D	E	F
1	Item#	Ability	Obj Prob	Sbj Prob	Over/Under	Weight
2	1	-2.16	0.095	0.600	0.505	0.025126
3	2	-2.16	0.034	0.533	0.499	0.025126
4	3	-2.16	0.024	0.533	0.509	0.025126
5	4	-2.16	0.191	0.467	0.276	0.025126
6	5	-2.16	0.272	0.667	0.395	0.025126

Figure 2. Calculating the weight statistics

Weighting the over/underconfidence scores

To apply the weighting statistic described above, simply add a column that multiplies the over/underconfidence score by the weighting score. This produces a metric of over/underconfidence that is weighted for the n size of examinees at each particular Rasch ability estimate level (θ_{ac}). By themselves, these numbers are not very informative, as they are only an intermediate step toward calculating the IDI. See Figure 3 for an example.

	A	B	C	D	E	F	G
1	Item#	Ability	Obj Prob	Sbj Prob	Over/Under	Weight	Over/Under*Weight
2	1	-2.16	0.095	0.600	0.505	0.025126	0.012688442
3	2	-2.16	0.034	0.533	0.499	0.025126	0.012546064
4	3	-2.16	0.024	0.533	0.509	0.025126	0.01279732
5	4	-2.16	0.191	0.467	0.276	0.025126	0.006926298
6	5	-2.16	0.272	0.667	0.395	0.025126	0.009916248

Figure 3. Calculating the weighted over/underconfidence scores

Finalizing the IDI calculations

To calculate the IDI for each item, one simply adds up all of the weighted over/underconfidence scores for each item. This provides a single statistic for each item that summarizes the overall amount of over/underconfidence on the item, controlling for frequency.

To quickly calculate this sum, the SUMIF function can be employed. For example, the following:

```
=SUMIF('IDI Calcs'!A:A,'Summary Stats'!A2,'IDI Calcs'!G:G)
```

instructs Excel to add up the weighted over/underconfidence scores from the IDI Calcs sheet for all rows whose item number matches that found in A2 of the Summary Stats sheet.

Interpreting the IDI

Paek et al. recommend using the differential item functioning (DIF) effect size scale recommended by Dorans and Holland (1993) to determine the size of the over/underconfidence. The scale is applied to the IDI as below:

$$\text{Large discrepancy} = IDI_i > |0.10| \tag{5}$$

$$\text{Medium discrepancy} = |0.05| < IDI_i \leq |0.10| \tag{6}$$

$$\text{Small or negligible discrepancy} = 0 < IDI_i \leq |0.05| \tag{7}$$

Therefore any IDI over 0.10 indicates a large degree of overconfidence, whereas one below -0.10 indicates a large degree of underconfidence. Medium discrepancies are found between 0.05 and 0.10, either positive or negative, and small discrepancies are those with IDIs under 0.05, either positive or negative.

Calculating the Discrepancy Percentage (DP)

Paek et al. describe two test-level summary statistics of over/underconfidence, but the first, the Test Discrepancy Index (TDI) is rather difficult to interpret, so they describe a transformation of the TDI to a simple percentage, called the DP.

The DP is expressed as follows (Paek et al., 2008, p. 5):

$$\text{Discrepancy Percentage (DP)} = \frac{\sum_{\theta_{ac}}[\sum_i P_i^* - \sum_i P_i]w(\theta_{ac})}{\text{Test Length}} \times 100 \tag{8}$$

Once again, the equation appears much more complex than the procedure actually is. In this case, we have already calculated the bulk of the numerator, as the second half of it is simply the weighted over/underconfidence scores. The rest of the numerator is simply adding all of them up for all of the items and all of the ability levels (θ_{ac}). If you have calculated them in a spreadsheet software package as described here; they are all in one column. This sum is the TDI, but we will transform it to a percentage to make it easier to interpret.

The sum of all the weighted over/underconfidence scores (aka the TDI) is divided by the test length (i.e., the number of items on the instrument), and this product is multiplied by 100 to transform it into a percentage. The result is the DP, which represents the percentage of over or underconfidence on the entire test.

Application of the method to the example dataset

Using the IDI and DP to investigate overconfidence at the item and test levels

The Paek et al. method described above was applied to the data described in the Data section. The IDIs and their effect sizes can be found in Table 1.

Table 1. ***IDIs and effect sizes for the items on the utterance verb instrument.***

Item	Word	IDI	Effect Size
1	speak	0.006	S
2	speak	0.260	L
3	speak	0.319	L
4	speak	-0.086	S
5	speak	-0.086	S
6	talk	0.255	L
7	talk	0.232	L
8	talk	0.225	L
9	talk	0.477	L
10	talk	0.318	L
11	say	0.014	S
12	say	0.013	S
13	say	0.071	M
14	say	-0.156	S
15	say	0.128	L
16	tell	0.075	M
17	tell	-0.034	S
18	tell	0.037	S
19	tell	0.001	S
20	tell	-0.017	S

The application of the Paek et al. method to these data reveals a large discrepancy between the respondents' accuracy on the items focusing on the use of the verb "talk". Since the direction of the discrepancy is positive, the IDIs are indicative of a high degree of overconfidence regarding the use of this verb. The verb "tell" seems to have little discrepancy between accuracy and confidence, and the other two verbs display a mix of over- and under confidence, as well as degree of discrepancy. The overall level of overconfidence as indicated by the Discrepancy Percentage (DP) was 10.26%, indicating that respondents were, on average, approximately 10% more confident of their answers than their actual accuracy.

Plotting the objective and subjective probabilities

Item-level plots

Another useful application of the Paek et al. method for investigating over/underconfidence on items is plotting the objective probabilities (P_i) against the subjective probabilities (P^*_i) for individual items. Doing so allows the researcher to investigate how the discrepancy changes with accuracy, and can shed a considerable amount of light upon the internal workings of examinees' minds at different levels of ability, as they encounter test items. See Figures 4 and 5 for examples.

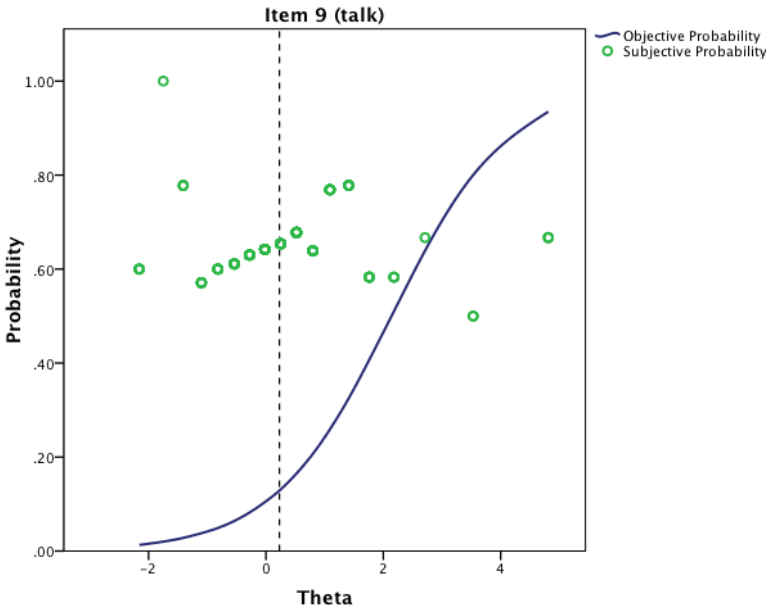

Figure 4. Plot of objective probability (accuracy) and subjective probability (confidence) against Rasch theta (ability/difficulty) for item 9. The dotted line represents the Rasch difficulty of the item.

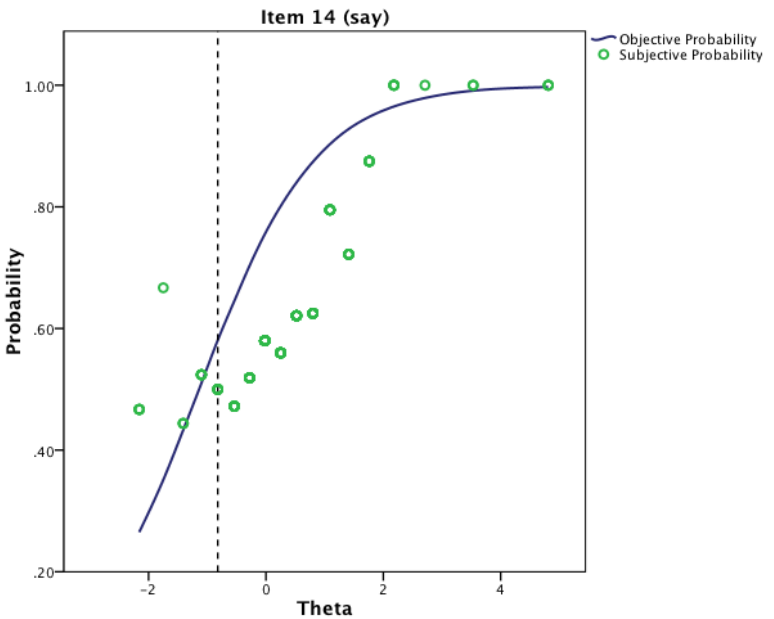

Figure 5. Plot of objective probability (accuracy) and subjective probability (confidence) against Rasch theta (ability/difficulty) for item 14. The dotted line represents the Rasch difficulty of the item.

Figure 4 is a plot of the objective probability (accuracy) and subjective probability (confidence) against the Rasch theta (the ability of the examinees and the difficulty of the items). The dotted line simply

serves as a reference point at the location of the difficulty of the item. In the case of item 9, which was the item with the highest positive IDI, indicating overconfidence, it is clear that lower-ability examinees are much more confident than their accuracy justifies. Interestingly, however, higher-ability examinees have similar confidence levels. This suggests that regardless of ability, examinees have roughly the same degree of confidence of their answers on this item. In the forthcoming paper on this study, we will discuss this further, but such a discussion is beyond the scope of this paper.

Figure 5 is a plot of item14, which was the item with the largest degree of underconfidence, as signified by the lowest IDI value. The relationship between confidence and accuracy here is interesting, in that it follows a roughly S-shaped curve, with low-ability examinees being slightly overconfident, mid-ability examinees being underconfident, and high-ability examinees' confidence and accuracy matching fairly closely.

To create the plots, a high-quality charting program is recommended. I have used the Simple Scatter/Dot chart type in SPSS 20 (IBM Corp., 2011). Both objective probability and subjective probability are assigned to the Y-axis, and the Rasch ability score is assigned to the X-axis. The objective probability plot is set to an interpolation line with the "Spline" setting, which makes interpreting the relationship between objective and subjective probabilities simpler. Finally, the optional line at the difficulty theta of the item is added manually via the "Reference Line" feature.

Other statistical or graphical packages are sure to include sufficient features to display these plots, but they are too involved for Microsoft Excel or other spreadsheet applications.

Test-level plots

Similar plots can be produced to examine the interaction between accuracy and confidence at the text level by adding up all the objective probabilities (P_i) and subjective probabilities (P^*_i) for each ability level (θ_{ac}). It is important to remind the reader that these probabilities must be from those calculated for the IDI calculations, so there is only one probability type per item, per ability level. Once again, the Excel SUMIF function is recommended here.

Once these are summed, they can be plotted against each other as in Figure 6. Once again, both the objective and subjective probabilities are assigned to the Y-axis, and the Rasch ability level (a.k.a. theta) to the X-axis.

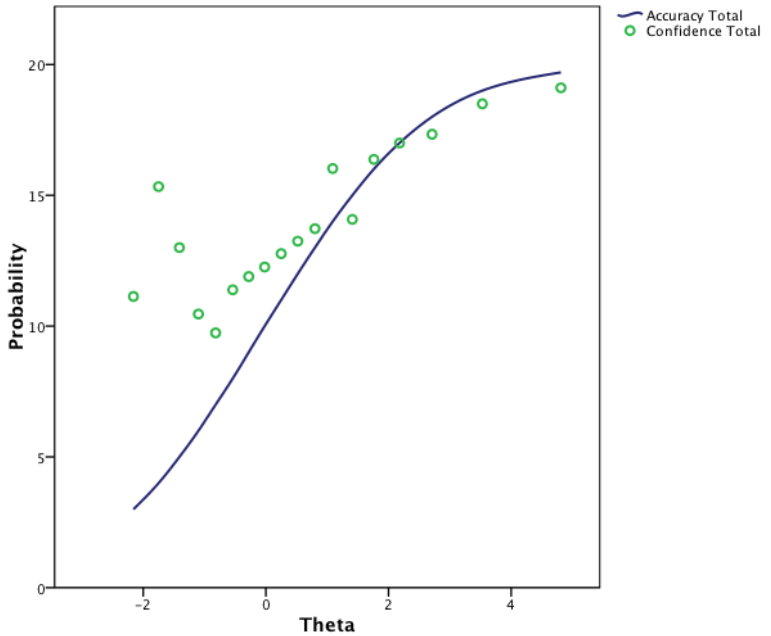

Figure 6. Plot of total objective probability (accuracy) and total subjective probability (confidence) against Rasch theta (ability/difficulty) for the entire test.

In the case of these data, it can be seen that lower-ability respondents tend to be overconfident, and higher-ability respondents tend to be underconfident. Once again, the implications of these findings within the scope of the study in question are beyond the scope of this paper, and will appear in a forthcoming paper by Sato and Batty.

Caveats regarding choice of model

Limitations of the Rasch model

One problem with the use of the Rasch model to examine confidence is that it assumes equal discrimination between items (i.e., all items discriminate between low- and high-ability examinees equally), typically with a slope of 1 (De Ayala, 2009). If one's items deviate from that assumption by a great deal, this presumption of equal discrimination can hide or otherwise mis-characterize the discrepancy between accuracy and confidence at the item level. For example, Figures 7 and 8 present the plots of probability against ability for items 9 and 14, originally discussed above, when scaled under the 2PL model instead of the Rasch model. Although 14 is similar to its Rasch counterpart, the addition of information about the discrimination of item 9 reveals the real source of the discrepancy: the item characteristic curve (ICC) is actually almost completely flat, meaning that examinees at all ability levels were roughly as likely to answer it correctly, so any interpretation of the discrepancy between accuracy and confidence here is essentially impossible. The item is simply far too hard, with a discrimination index of 0.082 and a difficulty measure of 18.905! In this case, test-level judgments based on the Rasch model may be valid, but item-level judgments would be very problematic.

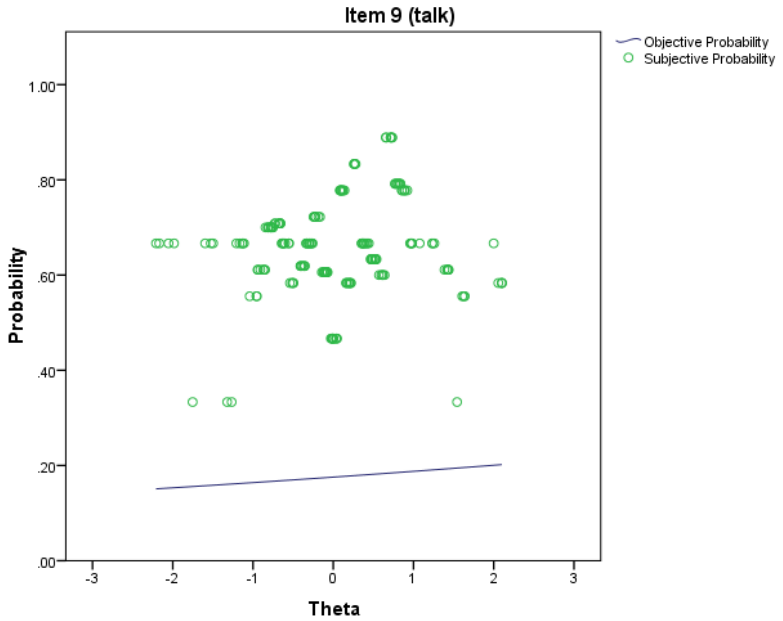

Figure 7. **Plot of objective probability (accuracy) and subjective probability (confidence) against 2PL theta (ability/difficulty) for item 9. The difficulty of the item is off the scale of the graph.**

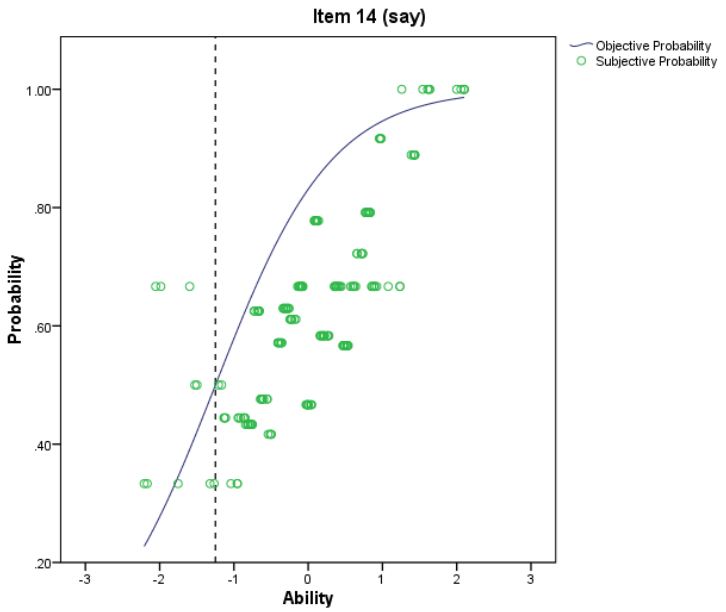

Figure 8. **Plot of objective probability (accuracy) and subjective probability (confidence) against 2PL theta (ability/difficulty) for item 14. The dotted line represents the Rasch difficulty of the item.**

Application of multi-parameter IRT models

If one concludes that a multi-parameter model would be more appropriate for the data at hand, the use of the two- or three-parameter logistic model with the Paek et al. method follows precisely the same steps as those for the Rasch model (see Stankov, Lee, & Paek, 2009). It may not be possible, however, to group respondents strictly by ability score for the calculation of subjective probability scores, as the addition of extra parameters results in fewer respondents with identical scores, rendering any kind of abstraction difficult. For the graphs presented in Figures 7 and 8, ability scores were grouped at the tenth-of-a-logit level (e.g., respondents with ability scores of 0.055 through 0.1444 would be grouped at the 0.1 level, by rounding to the nearest tenth of a logit) for the calculation of subjective probabilities and the weighting terms.

Conclusion

This paper has attempted to provide a more-accessible, practical explanation of the method developed by Paek et al. for examining the interaction between respondent confidence and accuracy using Rasch ability levels. Finally, the importance of adequately exploring one's data and considering its fit to the intended model was demonstrated.

I have focused my explanation here on only the most-straightforward of the analyses developed by Paek et al., as they are likely the most instructive and the most desired by most language-testing researchers. The original Paek et al. paper includes further analyses which require slightly more statistical knowledge to perform, but which are nonetheless very useful. Before applying the method discussed here, however, the reader is strongly encouraged to read the Paek et al. paper(s) and use the present article as a practical guide to applying the method to one's data.

Acknowledgements

This research was funded in part by Keio Gijuku Academic Development Funds (慶應義塾学事振興資金) at Keio University, Fujisawa, Japan.

References

De Ayala, R. J. (2009). *The theory and practice of item response theory*. New York: Guilford Press.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35–66). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hakstian, A. R., & Kansup, W. (1975). A Comparison of Several Methods of Assessing Partial Knowledge in Multiple-Choice Tests: Ii. Testing Procedures*. *Journal of Educational Measurement*, *12*(4), 231–239. doi:10.1111/j.1745-3984.1975.tb01024.x

IBM Corp. (2011). *IBM SPSS Statistics*. Armonk, NY: IBM Corp.

Keren, G. (1991). Calibration and probability judgements: Conceptual and methodological issues. *Acta Psychologica*, *77*(3), 217–273. doi:10.1016/0001-6918(91)90036-Y

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1981). *Calibration of Probabilities: The State of the Art to 1980*.

Linacre, J. M. (2012a). Winsteps (Version 3.75.0) [Computer Software]. Beaverton, Oregon: Winsteps.com. Retrieved from http://www.winsteps.com/

Linacre, J. M. (2012b). Facets (Version 3.70.0) [Computer Software]. Beaverton, Oregon: Winsteps.com. Retrieved from http://www.winsteps.com/facets.htm

Paek, I., Lee, J., Stankov, L., & Wilson, M. (2008). *A study of confidence and accuracy using the Rasch modeling procedures* (Research Report No. RR-08-42). Princeton, NJ: Educational Testing Service.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen, Denmark: Danmarks Paedogogiske Institut.

Sato, Y., & Batty, A. (2012). A study of learners' intuitions behind the use of utterance verbs in English. *Vocabulary Learning and Instruction*, *1*(1), 29–36. doi:10.7820/vli.v01.1.sato.batty

Stankov, L., Lee, J., & Paek, I. (2009). Realism of confidence judgments. *European Journal of Psychological Assessment*, *25*(2), 123–130. doi:10.1027/1015-5759.25.2.123