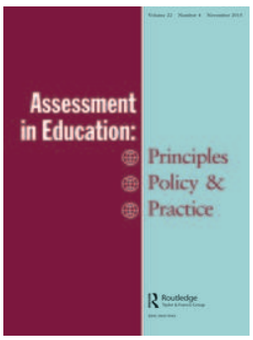

Routledge
Taylor & Francis Group

Assessment in Education: Principles, Policy & Practice

ISSN: 0969-594X (Print) 1465-329X (Online) Journal homepage: http://www.tandfonline.com/loi/caie20

Accuracy in student self-assessment: directions and cautions for research

Gavin T.L. Brown, Heidi L. Andrade & Fei Chen

To cite this article: Gavin T.L. Brown, Heidi L. Andrade & Fei Chen (2015) Accuracy in student self-assessment: directions and cautions for research, Assessment in Education: Principles, Policy & Practice, 22:4, 444-457, DOI: 10.1080/0969594X.2014.996523

To link to this article: http://dx.doi.org/10.1080/0969594X.2014.996523

Assessment in Education: Principles, Policy & Practice, 2015
Vol. 22, No. 4, 444–457, http://dx.doi.org/10.1080/0969594X.2014.996523

Routledge
Taylor & Francis Group

Accuracy in student self-assessment: directions and cautions for research

Gavin T.L. Brown[a]*, Heidi L. Andrade[b] and Fei Chen[b]

[a]*School of Learning, Development and Professional Practice, Faculty of Education, The University of Auckland, Auckland, New Zealand;* [b]*Educational Psychology and Methodology, The State University of New York, Albany, NY, USA*

(*Received 5 November 2013; accepted 25 November 2014*)

Student self-assessment is a central component of current conceptions of formative and classroom assessment. The research on self-assessment has focused on its efficacy in promoting both academic achievement and self-regulated learning, with little concern for issues of validity. Because reliability of testing is considered a sine qua non for the validity of assessment interpretations, and research into the human ability to self-evaluate work raises concerns about the quality of students' judgements, it is sensible to investigate the accuracy of students' self-assessments. This article reviews relevant literature from educational psychology and psychometrics to define the need for a better understanding of accuracy in self-assessment as well as to identify possible pitfalls in measuring accuracy that could undermine its effectiveness by, for example, trading the focus on formative feedback for summative scoring or rating. The article concludes with recommendations for the design of research on accuracy in self-assessment.

Keywords: self-assessment; accuracy; literature review; research implications

Self-assessment in education involves reflecting on and monitoring one's own work processes and/or products (Brown & Harris, 2013). Self-assessment can involve both description (i.e. these are the characteristics of my work) and evaluation (i.e. this is how good my work is and what it is worth). Student self-assessment is a central component of current conceptions of classroom assessment, particularly formative assessment (Andrade, 2010; Leahy, Lyon, Thompson, & Wiliam, 2005), and a number of studies have demonstrated a positive association between self-assessment, learning and achievement (see Brown & Harris, 2013, for a review). However, the research on self-assessment has focused predominantly on its efficacy in promoting both academic achievement and self-regulated learning, with little concern for issues of validity.

Validity requires that inferences drawn from an assessment lead to appropriate interpretations and actions (Kane, 2006; Messick, 1989). In order to support valid uses, an assessment is expected to be an accurate and dependable portrayal of a learner's achievement or performance. The question for self-assessment is to what degree student self-descriptions and evaluations of their work are truthful or veridical, as Butler (2011) put it.

*Corresponding author. Email: gt.brown@auckland.ac.nz

However, the concept of accuracy in self-assessment is thorny. Within classical test theory, the true score of a performance or product is estimated by determining the degree of inconsistency in a score (Haertel, 2006). Item response theory does not obviate this problem; error is still present, though perhaps better handled (Embretson & Reise, 2000). While decontextualised, absolute objectivity is not possible, we can argue for veridicality in self-assessment by comparing those results with other measures of achievement or performance. Yet consistency between teacher and student, peer and student, or test and student might arise from socially shared, systematic errors (e.g. bias to avoid low grades or scores). Nonetheless, there is widespread agreement that where many judges and measures agree about the quality of a product or process, their combined evaluation must be in some sense 'true' (Brennan, 2001). Hence, accuracy in self-assessment can be determined by comparing a student's self-assessment to the judgements of qualified raters, such as teachers or fellow learners, or to performance on tests.

Although there is evidence that students can be consistent with their own previous self-assessments (Ross, 2006), Brown and Harris (2013) found that the correlation between self-ratings and teacher ratings, between self-estimates of performance on a test and actual test scores, and between student and teacher judgements based on a rubric tended to be only weakly positive. Further, because self-assessments take place in many different ways (e.g. estimating total score before or after taking a test, deciding if a test item was answered correctly, estimating how many tasks can be completed in a set time, self-assessment according to a rubric, ranking one's work relative to that of others, etc.), how well students can self-assess with one method may have no relationship to how well they use another (Bol & Hacker, 2012; Maki, Shields, Wheeler, & Zacchilli, 2005).

Dunning, Heath, and Suls (2004) identified many reasons by which self-assessments can be inaccurate, including the tendencies to (a) be unrealistically optimistic about one's own abilities, (b) believe that one is above average, (c) neglect crucial information and (d) have deficits in information. Inaccurate self-assessment might also be attributed to the social environment of classrooms, where the pressure to enhance or even protect one's own self-worth can result in overestimation of one's ability (Saavedra & Kwun, 1993), or inaccurate self-reporting of grades or test scores (Kuncel, Credé, & Thomas, 2005). Students have also been found to consider their own effort, which ought to be independent of a quality evaluation of work products, when assessing their work (Ross, Rolheiser, & Hogaboam-Grey, 1998b). While some inaccuracy may be deliberate, much inaccuracy may be unintentional or subconscious.

Inconsistency in or lack of skill in self-assessment is not evidence in itself that inaccuracy in self-assessment, however it is understood, is problematic. So the question facing us is simple: Does it matter if students are inaccurate in their self-assessments, so long as they are engaged in thinking about the quality of their work? It seems logical that students who conclude wrongly that they are good or weak in some domain will base personal decisions on such false interpretations, possibly bringing about harm to their learning (e.g. task avoidance and not enrolling in future subjects) (Schunk & Pajares, 2004). This hypothesis is in need of rigorous testing, however, and the testing itself could be fraught with problems. We will discuss the potential pitfalls after a brief review of the literature on accuracy in self-assessment.

Review of literature on accuracy in self-assessment

The correlation between student self-ratings and other measures tended to be positive, ranging from weak to moderate (i.e. values ranging from $r \approx .20$ to .80), with few studies reporting correlations greater than $r = .60$ (Brown & Harris, 2013). A striking characteristic of more accurate self-assessors is that they tend to be less optimistic than more inaccurate self-assessors: in other words, greater competence with self-assessment is associated with more humble self-evaluation (Blatchford, 1997a, 1997b; Eccles, Wigfield, Harold, & Blumenfeld, 1993; Frey & Ruble, 1987; Kaderavek, Gillam, Ukrainetz, Justice, & Eisenberg, 2004; Kasanen, Räty, & Eklund, 2009; Ross, Rolheiser, & Hogaboam-Grey, 2002; Stipek, 1981; Stipek & Tannatt, 1984; Wilson & Wright, 1993).

Accuracy, in terms of consistency with other measures, seems to be higher for test scores than for global qualities like cognitive competence or for complex performances such as writing and projects. Studies of consistency tend to fall into two categories: student self-assessments relative to teacher ratings and self-assessments relative to test scores.

Studies of self-assessments and teacher ratings

A study of American elementary students (Connell & Ilardi, 1987) in Grades 4–6 found a moderate correlation ($r = .41$) between student self- and teacher ratings of cognitive competence. Using student-created criteria on a five-point scale, 54% of American Grade 1 and 2 student self-evaluations of social studies group projects matched exactly the teacher ratings (Higgins, Harris, & Kuehn, 1994). An experimental study of American Grade 7 science students involved students in creating a rubric for scoring tests and found that, although the correlation between student self-grading and teacher grading was very high ($r = .98$), 100% of the 24 students in the self-grading condition overgraded their performance relative to teacher grading (Sadler & Good, 2006). In contrast, Sung, Chang, Chang, and Yu (2010) reported two studies of self-assessment with Taiwanese middle schoolers (i.e. Grade 7 recorder playing and Grade 8 group projects to create a multimedia web page) and found much weaker correlations between students and teachers (mean correlation $r = .41$ for recorder playing; $r = .52$ for group project performance). Similarly, Butler (1990) reported that the Israeli elementary school students (Grades K, 2, and 5) in her study consistently over-rated the quality of their drawings relative to teachers, especially under conditions of competition.

Studies of self-assessments and test scores

Except for a few studies, there seems to be reasonably strong evidence that students can judge how well they have done on formal tests and assessments. The exceptions include a study of elementary grade, low socio-economic, American students (Bradshaw, 2001), in which the correlation was just $r = .26$ between actual performance and self-estimate of performance immediately after a test of reading comprehension. Similarly, amongst 311 American children in Grades 4–6, self-assessment of their performance, using a five-point scale, immediately after completing five standardised achievement tests was relatively weak (i.e. mean $r = .27$, SD = .04) (LaVoie & Hodapp, 1987). Likewise, a survey study of 365 students who earned

entrance to the 11th year of schooling by passing the Bhutan Certificate of Secondary Education external examination resulted in a low to moderate correlation ($r = .30$) between student self-rating and actual performance (Luyten & Dolkar, 2010).

In contrast, Wilson and Wright (1993) reported that 300 rural American high school students (Grades 8–12, with 49% in Grade 11) estimated their own scores on standardised numerical and verbal ability tests, and the self-assessments correlated moderately with their actual performance (mean $r = .48$, SD = .13). A large-scale survey study (Alsaker, 1989) found moderately strong correlations (ranging from $r = .58$ to .72) between self-ratings of global academic competence and average test score for Norwegian, English and mathematics. A small survey study of Japanese high school and university students found moderately strong correlations between self-appraised ability in language ($r = .66$) and speaking ($r = .65$), and actual tested performance (Ikeguchi, 1996).

From this sample of studies, we can see generally weak to moderate levels of agreement between student self-assessments and the ratings of their teachers or performances on tests. Nonetheless, there appear to be factors which mitigate the similarity of assessments, which we will discuss next.

Factors related to consistency

Consistency seems to improve with age and experience with school. For example, Blatchford's (1997a) longitudinal study in the UK found that the average self-assessed grade decreased from age 7 to 16, but the accuracy of the student-assigned grade relative to actual performance increased in the same period. Likewise, Hewitt (2005) found that the relationship between student self-evaluation of musical performance and the ratings of expert music educators was more consistent amongst high school students than middle school students across all subareas of musical performance by statistically significant margins. Similarly, accuracy in estimating reading comprehension test performance was better for Grade 5 American students than those in either Grade 3 or 4 (Bradshaw, 2001).

A study of American elementary students (ages 5–12) (Kaderavek et al., 2004) found that the proportion of students who did poorly and who overestimated their performance declined from 85% at ages 5–9 to just 43% of 10- and 11-year olds ($d = 1.21$). Alsaker (1989) reported that the correlation between self-perceived global academic competence and average test scores increased with higher grades and suggested this may be due to the 'older children's higher level of cognitive and emotional development' (p. 156). Butler (1990) found that, compared to Kindergarten students ($d = 1.18$, mean $r = .32$), Grade 5 Israeli student self-evaluations of their drawings were more like those of their teachers ($d = -.07$, mean $r = .70$).

Greater academic ability or performance is also associated with a stronger tendency to self-evaluate in a manner more consistent with teachers and test scores (Barnett & Hixon, 1997). For example, a large survey of Canadian middle school students found that those who could use a rubric to score work in the same fashion as the official scoring rubric tended to perform better in writing (Laveault & Miles, 2002). They also tended to be more severe in their marking. Birnbaum (1972) tracked nearly 8000 high school graduates and found that there was a clear tendency for students with higher actual averages to be more accurate, or possibly more honest, in reporting their high school grade average than those with lower averages.

An experimental study of French-speaking students in the first two grades of secondary schooling in Montreal found that the accuracy of the self-ratings of high-performing students was higher than those of the low-performing students by a statistically significant margin ($d = .84$; Claes & Salame, 1975). However, in an Australian study of 94 final-year (average age 17 years, 8 months) high school students, Ng and Earl (2008) found that students whose estimates of proficiency in a school-based trial examination of English were *under*-estimates relative to their actual score had higher scores by statistically significant margins than those who made accurate or over-estimations of their actual performance, suggesting that consistency in self-assessment is not completely assured even for high-performing students.

Aspects of the product or performance being evaluated also contribute to the veridicality of student self-assessments. The more simple and concrete the task (Barnett & Hixon, 1997; Bradshaw, 2001; Hewitt, 2005), and the more specific and concrete the reference criteria (Claes & Salame, 1975), the more accurately students estimated their own performance. Comparison to explicitly stated criteria, goals or standards as the basis for self-assessment can also improve the veridicality of self-assessments (Andrade & Valtcheva, 2009). Accuracy was improved when a small sample of Grades 5 and 6 American students were explicitly taught to use a self-checking strategy when solving long division problems (Ramdass & Zimmerman, 2008).

Thus, the accuracy of student self-assessment does not appear to be uniform throughout the student's life course, nor across the full range of learning activities. In general, these findings reflect the results of research in the distinct but related field of calibration, which is the degree of fit between a person's judgement of performance and his or her actual performance (Bol & Hacker, 2012). Research on calibration suggests that rewards can increase accuracy (Miller, Duffy, & Zane, 1993), as can feedback. For example, Miller and Geraci (2011) examined the relation between the accuracy of 81 undergraduate students' predictions of their performance on four exams and (1) extra-credit incentives and (2) explicit, concrete feedback. They found that the incentives and feedback were related to improved accuracy for low-performing students but, interestingly, not to higher scores on exams. Predictions about performance on exams are only one type of self-assessment, however. Little is known about whether or not training in self-assessment influences accuracy in a classroom context where self-assessment is used formatively to guide revision and improvement.

Implications for research

An obvious implication of the foregoing discussion is that we need to know more about the nature and nurture of accuracy in classroom-based self-assessment. Less obvious implications are related to potential pitfalls in examining accuracy, which are discussed below. Here we summarise what we know about designing and implementing effective self-assessment, if not actually improving accuracy. We recommend that researchers integrate the following features into their research on accuracy in self-assessment.

Clear criteria

Sharing learning targets and criteria is generally considered good assessment practice (Brookhart, 2013), and Falchikov and Boud's (1989) meta-analysis found

student familiarity with rating criteria enhanced accuracy and the alignment of ratings with academics. Indeed, having students involved in the process of creating criteria for rubrics is especially associated with greater learning outcomes (Andrade, Du, & Mycek, 2010; Andrade, Du, & Wang, 2008; Ross, Rolheiser, & Hogaboam-Grey, 1998a; Sadler & Good, 2006).

Models

Samples of target performances, particularly exemplars, are associated with improved performance (Andrade et al., 2010; Hewitt, 2001), and might enhance accuracy if the models are used as benchmarks (Dunning et al., 2004).

Instruction and practice

As with all learning, student self-assessment needs to be taught (Brown & Harris, 2014; Ross, 2006). Epley and Gilovich (2005) claim that when people are overconfident in the accuracy of their judgements, it is because they think too little about the ways in which they might be wrong: errors in judgement are reduced when people 'pause for a moment and think a bit harder' (p. 200). Research on the effects of practice on accuracy is mixed – e.g. Brookhart, Andolina, Zuza, and Furman (2004) and Lopez and Kossack (2007) found student self-assessments became more realistic with practice, while Lew, Alwis, and Schmidt (2010) found that accuracy in assessment did not improve over time.

Not surprisingly, it appears that practice alone is an insufficient condition. This conclusion is echoed in the calibration literature, which has demonstrated that feedback and practice alone tend to be insufficient for debiasing calibration by low-achieving students (Hacker, Bol, Horgan, & Rakow, 2000; Nietfeld, Cao, & Osborne, 2005). The types of instruction and practice that show promise in improving calibration accuracy include reflection, instruction in or guidelines for monitoring, feedback on initial adjustments, incentives for improved accuracy (for low-achieving students) and group calibration practice during which students discuss and evaluate their own understandings with their peers (Bol & Hacker, 2012).

Feedback on accuracy

Feedback from others as to the accuracy of students' predictions of performance has the potential to improve accuracy (Miller & Geraci, 2011). Self-monitoring and reflection might also encourage students to compare their self-assessment and actual performance over time, thereby enabling more accurate self-assessment (Lopez & Kossack, 2007).

Rewards

Getting students to set stringent targets for self-selected rewards combined with self-monitoring or self-marking has led to improved learning outcomes (Barling, 1980; Miller et al., 1993; Wall, 1982). However, the effects of incentives on accuracy tend to be mixed (Hacker, Bol, & Bahbahani, 2008), so this tactic should be employed with caution.

Keep it formative

Including student self-assessments as part of summative course grades introduces high-stakes consequences for honest, accurate evaluations (Andrade, 2010; Brown & Harris, 2013). We recommend studying formative self-assessment in contexts that do not tempt students to inflate or distort their self-evaluations.

Pitfalls to avoid when investigating accuracy

The review of the literature reveals that our understanding of the nature and role of accuracy in self-assessment is incomplete. This gap in our knowledge implies that teachers and researchers must collect data on the correlations between student self-assessments and teacher assessments, and, more importantly, teach inaccurate assessors (whether high or low performers) how to produce more valid and realistic evaluations of their work. On closer inspection, however, such an approach is shown to be somewhat complicated. With concerns for consequential validity (Messick, 1989), in this section, we consider several methodological pitfalls – reliability, grading, social response bias, response style and trust/respect – and recommend possible ways to avoid them.

The reliability pitfall

Comparing students' self-assessment to teachers' evaluations of their work can be problematic, because of the notorious lack of accuracy/reliability in teacher grading (Heldsinger & Humphry, 2013; Kirby & Downs, 2007), even when teachers have a common understanding of the criteria for a piece of student work (Falchikov, 2005). Kirby and Downs (2007) found that inaccuracies and leniency in teachers' grading influenced estimates of students' accuracy. Consequently, Kirby and Downs (2007) urge us to avoid assuming that agreement is an indicator of accurate student self-assessment: close examination of the quality of teachers' (or any raters') evaluations is important.

The grading pitfall

In a study of undergraduate writers, Lipnevich and Smith (2008) found that students who were shown the grade they received for their first draft performed less well on the final version than those who were not shown their grade. Their study predicts a performance decrement if students are shown their teacher's evaluation of their work while it is still in progress. Studies of accuracy would need to avoid inadvertently recreating this phenomenon by implying to students that a teacher's assessment of their formative drafts is a summative or final score or grade, or even predictive of the terminal evaluation score.

The social response bias pitfall

We have long known that the validity of self-report measures, especially those that involve public disclosure, is threatened by socially desirable responding (Paulhus, 1991), particularly for females (Dalton & Ortegren, 2011). A method that would be useful to researchers (but perhaps not feasible for teachers) is to assure students of

anonymity by asking them to hand in assignments and self-assessments with no identifying information attached.

Researchers who cannot control for social response bias can measure and account for it. For example, a measure of socially desirable responding that can detect faking (e.g. scales that reveal degrees of self-deceptive positivity or impression management) could be administered along with the self-assessment (Paulhus, 1991) in order to allow degree of bias to be statistically identified and controlled in subsequent analyses of self-assessments. Paulhus (1991) also recommends minimising socially desirable responding by giving students sufficient time to engage in a task (in this case, self-assessment), avoiding emotional arousal by not associating the task with high stakes and minimising distractions.

The response style pitfall

Consistent response styles exist in how members of various groups evaluate their own identities or worth; some students may be consistently negative or positive in their self-perceptions. As Kasanen and Räty (2002) make clear, self-assessment is somewhat of a misnomer, since it is not an evaluation of the self. Thus, self-assessment researchers have to help students to focus on the quality of their work, as an independent artefact, rather than as an extension of their ego. Such a focus would mean students would be allowed to keep their ego intact, while giving their work a relatively negative, but realistic, evaluation. In addition to making an effort to remove the ego from self-assessment, researchers could include a measure of response style in their research design.

The classroom environment pitfall: trust and respect

Students have differing and highly personal reactions to self-assessment disclosure (Cowie, 2009; Harris, Harnett, & Brown, 2009). Some have raised concerns about their psychological safety when their self-evaluations are made public to peers, parents and teachers (Cowie, 2009; Harris & Brown, 2013; Raider-Roth, 2005; Ross et al., 1998b, 2002). Some students provide depressed self-evaluations for fear of being seen as egotistical (Brooks, 2002) or because of cultural practices such as self-effacement (Kwok & Lai, 1993). Others give elevated self-assessments to avoid being shamed in front of the class (Harris & Brown, 2013). While peer feedback may help students calibrate their self-assessment and, thereby, increase accuracy (Dunning et al., 2004), there are potential pitfalls in students' interpersonal relationships that can undermine both self-assessment and peer response (Brown & Harris, 2013; Harris & Brown, 2013).

Noting that the self-assessment work she studied in a sixth-grade classroom was 'embedded in the web of classroom relationships and, as such, was a deeply relational process', Raider-Roth (2005, p. 9) reported that students carefully selected what they would disclose to teachers. For example, one girl told of not admitting to being good at writing paragraphs, because she did not want the teacher to talk about it in front of the whole room of students, while another stressed the need to get the self-assessment work 'right', meaning what the teacher expected.

Thus, it cannot be assumed that students are honest in self-assessments that are shared with others, nor that inaccurate self-assessments indicate a true misunderstanding on the part of the student. For example, one girl noted that although 'she

might selectively tell all the truth to the outside world, she does not avoid confronting the whole truth inside herself' (Raider-Roth, 2005, p. 128). If such attitudes and behaviours are common, shared self-assessments could provide counterfeit data. Because disclosure is highly individualised, and trust and respect are essential qualities of a classroom in which students are willing to disclose their knowledge and engage in assessment for learning (Tierney, 2010), researchers interested in investigating issues of accuracy in self-assessment will have to attend to the classroom environment and its effects on their data (Brown & Harris, 2013). Here again, assuring students of anonymity by asking them to hand in assignments and self-assessments with no identifying information attached might help researchers, if not teachers. Alternatively, researchers might allow students to restrict access to their self-assessments to those whom the student trusts and gives permission, potentially excluding the teacher (Cowie, 2009).

Conclusion

The pitfalls discussed above are complex and not easily resolved. Until we have a better understanding of the nature of self-assessment, the best we can do as researchers is to attempt to create the optimal conditions for accuracy and avoid known pitfalls, which include issues of reliability, grading, social response bias, response style and trust/respect. Reliability should be addressed by maximising the psychometric quality of any criterion used to judge the accuracy of student self-assessments (e.g. test, teacher rating, etc.): there is no point evaluating students' accuracy with an inaccurate measure. Problems related to grading and trust/respect can be managed by implementing self-assessment in a context likely to promote accuracy – or at least *not* promote inaccuracy – meaning that self-assessments should not count towards grades, and should be private. Social response bias and response style can be managed to some degree by encouraging students to be honest and accurate, but students' tendencies towards bias could also be measured.

Of course, we also recommend the use of randomised trials in order to establish causal inferences about the impact of variables on the accuracy of self-assessment. To date, most studies of self-assessment accuracy have not employed randomisation of assignment. Experimental studies that involve random assignment generally involve learners in tasks that lack authenticity and, thereby, limit the generalisability of results to classroom contexts. With large enough samples, it would be possible to statistically control for factors influencing accuracy, including prior ability or achievement level, task difficulty, assessment purpose and so on.

Given the social and interpersonal nature of student self-assessment in classroom contexts and the great variation in culturally preferred classroom climates and practices, more work needs to be done to understand whether the recommendations we have made are equally applicable in all societies. It is possible in Confucian-heritage cultures, for example, that the importance placed on high performance and the need to avoid low ranks, combined with the pressure from teachers and parents to continually do better (Brown & Wang, 2013), would discourage or even prevent realistic evaluations. In systems that are highly selective, it is difficult to expect the weakest students in a highly proficient class to realise that their work is still actually good. Likewise, students in systems or schools that provide relatively positive and inflated reports of proficiency (e.g. Hattie & Peddie's [2003] description of New Zealand primary school report cards) are unlikely to develop a realistic sense of the quality of

their work. The cultural context of self-assessment must be understood by researchers.

Finally, we recommend that researchers carefully track and report their efforts to optimise students' honest, insightful and evidence-based judgements. The outlines of a curriculum for the structured implementation of self-assessment in schooling are available (Brown & Harris, 2014), but considerable work is needed to refine those ideas. There is still much to understand concerning the accuracy of student self-assessment, as well as the effects of well-meaning attempts to improve it.

Notes on contributors

Gavin T.L. Brown is an associate professor of Education and Director of the Quantitative Data Analysis and Research unit in the Faculty of Education at the University of Auckland, New Zealand. He has published over 90 research articles in refereed journals and book chapters, and has written two textbooks on assessment and co-authored two standardised educational test systems. He is co-author of a recent review of student self-assessment published in *The Sage Handbook of Research on Classroom Assessment* (2013). His major research interest is cross-cultural study of the social psychological effects of assessment on prospective and in-service teachers and upon school and higher education students.

Heidi L. Andrade is an associate professor in Educational Psychology and Methodology, and the associate dean of academic affairs at the State University of New York, Albany. Her research and teaching focus on assessment and self-regulated learning, with an emphasis on student self-assessment. She has designed thinking-centred instruction and assessments for classrooms, after-school programmes, children's television shows, and CD-ROMs. She is an associate editor of and contributor to *The Sage Handbook of Research on Classroom Assessment* (2013).

Fei Chen is a PhD student in the educational psychology and methodology programme at the University at Albany, State University of New York, supervised by Dr Heidi Andrade. Her research interests include self-regulated learning, learning from instruction and assessment, and gifted education.

References

Alsaker, F. D. (1989). School achievement, perceived academic competence and global self-esteem. *School Psychology International, 10*, 147–158. doi:10.1177/0143034389102009

Andrade, H. (2010). Students as the definitive source of formative assessment: Academic self-assessment and the self-regulation of learning. In H. Andrade & G. Cizek (Eds.), *Handbook of formative assessment* (pp. 90–105). New York, NY: Routledge.

Andrade, H. L., Du, Y., & Mycek, K. (2010). Rubric-referenced self-assessment and middle school students' writing. *Assessment in Education: Principles, Policy & Practice, 17*, 199–214. doi:10.1080/09695941003696172

Andrade, H. L., Du, Y., & Wang, X. (2008). Putting rubrics to the test: The effect of a model, criteria generation, and rubric-referenced self-assessment on elementary school students' writing. *Educational Measurement: Issues and Practice, 27*, 3–13. doi:10.1111/j.1745-3992.2008.00118.x

Andrade, H., & Valtcheva, A. (2009). Promoting learning and achievement through self-assessment. *Theory into Practice, 48*, 12–19.

Barling, J. (1980). A multistage multidependent variable assessment of children's self-regulation of academic performance. *Child Behavior Therapy, 2*, 43–54.

Barnett, J. E., & Hixon, J. E. (1997). Effects of grade level and subject on student test score predictions. *The Journal of Educational Research, 90*, 170–174.

Birnbaum, R. (1972). Factors associated with the accuracy of self-reported high-school grades. *Psychology in the Schools, 9*, 364–370. doi:10.1002/1520-6807%28197210%299:4%3C364:AID-PITS2310090404%3E3.0.CO;2-G

Blatchford, P. (1997a). Pupils' self assessments of academic attainment at 7, 11 and 16 years: Effects of sex and ethnic group. *British Journal of Educational Psychology, 67*, 169–184.
Blatchford, P. (1997b). Students' self assessment of academic attainment: Accuracy and stability from 7 to 16 years and influence of domain and social comparison group. *Educational Psychology, 17*, 345–359. doi:10.1080/0144341970170308
Bol, L., & Hacker, D. J. (2012). Calibration research: Where do we go from here? *Frontiers in Psychology, 3*, doi:10.3389/fpsyg.2012.00229
Bradshaw, B. K. (2001). Do students effectively monitor their comprehension? *Reading Horizons, 41*, 143–154.
Brennan, R. L. (2001). *Generalizability theory*. New York, NY: Springer.
Brookhart, S. M. (2013). *How to create and use rubrics for formative assessment and grading*. Alexandria, VA: Association for Supervision & Curriculum Development.
Brookhart, S. M., Andolina, M., Zuza, M., & Furman, R. (2004). Minute math: An action research study of student self-assessment. *Educational Studies in Mathematics, 57*, 213–227.
Brooks, V. (2002). *Assessment in secondary schools: The new teacher's guide to monitoring, assessment, recording, reporting and accountability*. Buckingham: Open University Press.
Brown, G. T. L., & Harris, L. R. (2013). Student self-assessment. In J. H. McMillan (Ed.), *The Sage handbook of research on classroom assessment* (pp. 367–393). Thousand Oaks, CA: Sage.
Brown, G. T. L., & Harris, L. R. (2014). The future of self-assessment in classroom practice: Reframing self-assessment as a core competency. *Frontline Learning Research, 2*, 22–30. doi:10.14786/flr.v2i1.24
Brown, G. T. L., & Wang, Z. (2013). Illustrating assessment: How Hong Kong university students conceive of the purposes of assessment. *Studies in Higher Education, 38*, 1037–1057. doi:10.1080/03075079.2011.616955
Butler, R. (1990). The effects of mastery and competitive conditions on self-assessment at different ages. *Child Development, 61*, 201–210.
Butler, R. (2011). Are positive illusions about academic competence always adaptive, under all circumstances: New results and future directions. *International Journal of Educational Research, 50*, 251–256. doi:10.1016/j.ijer.2011.08.006
Claes, M., & Salame, R. (1975). Motivation toward accomplishment and the self-evaluation of performances in relation to school achievement. *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement, 7*, 397–410. doi:10.1037/h0081924
Connell, J. P., & Ilardi, B. C. (1987). Self-system concomitants of discrepancies between children's and teachers' evaluations of academic competence. *Child Development, 58*, 1297–1307. doi:10.2307/1130622
Cowie, B. (2009). My teacher and my friends helped me learn: Student perceptions and experiences of classroom assessment. In D. M. McInerney, G. T. L. Brown, & G. A. D. Liem (Eds.), *Student perspectives on assessment: What students can tell us about assessment for learning* (pp. 85–105). Charlotte, NC: Information Age.
Dalton, D., & Ortegren, M. (2011). Gender differences in ethics research: The importance of controlling for the social desirability response bias. *Journal of Business Ethics, 103*, 73–93. doi:10.1007/s10551-011-0843-8
Dunning, D., Heath, C., & Suls, J. M. (2004). Flawed self-assessment: Implications for health, education, and the workplace. *Psychological Science in the Public Interest, 5*, 69–106.
Eccles, J., Wigfield, A., Harold, R. D., & Blumenfeld, P. (1993). Age and gender differences in children's self- and task perceptions during elementary school. *Child Development, 64*, 830–847.
Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.
Epley, N., & Gilovich, T. (2005). When effortful thinking influences judgmental anchoring: Differential effects of forewarning and incentives on self-generated and externally provided anchors. *Journal of Behavioral Decision Making, 18*, 199–212.

Falchikov, N. (2005). *Improving assessment through student involvement: Practical solutions for aiding learning in higher and further education*. London: Routledge Falmer.
Falchikov, N., & Boud, D. (1989). Student self-assessment in higher education: A meta-analysis. *Review of Educational Research, 59*, 395–430.
Frey, K. S., & Ruble, D. N. (1987). What children say about classroom performance: Sex and grade differences in perceived competence. *Child Development, 58*, 1066–1078.
Hacker, D. J., Bol, L., & Bahbahani, K. (2008). Explaining calibration accuracy in classroom contexts: The effects of incentives, reflection, and explanatory style. *Metacognition and Learning, 3*, 101–121.
Hacker, D., Bol, L., Horgan, D., & Rakow, E. (2000). Test prediction and performance in a classroom context. *Journal of Educational Psychology, 92*, 160–170.
Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 65–110). Westport, CT: Praeger.
Harris, L. R., & Brown, G. T. L. (2013). Opportunities and obstacles to consider when using peer- and self-assessment to improve student learning: Case studies into teachers' implementation. *Teaching and Teacher Education, 36*, 101–111. doi:10.1016/j.tate.2013.07.008
Harris, L. R., Harnett, J., & Brown, G. T. L. (2009). "Drawing" out student conceptions of assessment: Using pupils' pictures to examine their conceptions of assessment. In D. M. McInerney, G. T. L. Brown, & G. A. D. Liem (Eds.), *Student perspectives on assessment: What students can tell us about assessment for learning* (pp. 53–83). Charlotte, NC: Information Age.
Hattie, J., & Peddie, R. (2003). School reports: 'Praising with faint damns'. *Set: Research Information for Teachers,* (3), 4–9.
Heldsinger, S. A., & Humphry, S. M. (2013). Using calibrated exemplars in the teacher-assessment of writing: An empirical study. *Educational Research, 55*, 219–235.
Hewitt, M. P. (2001). The effects of modeling, self-evaluation, and self-listening on junior high instrumentalists' music performance and practice attitude. *Journal of Research in Music Education, 49*, 307–322. doi:10.2307/3345614
Hewitt, M. P. (2005). Self-evaluation accuracy among high school and middle school instrumentalists. *Journal of Research in Music Education, 53*, 148–161.
Higgins, K. M., Harris, N. A., & Kuehn, L. L. (1994). Placing assessment into the hands of young children: A study of student-generated criteria and self-assessment. *Educational Assessment, 2*, 309–324. doi:10.1207/s15326977ea0204_3
Ikeguchi, C. B. (1996). *Self assessment and ESL competence of Japanese returnees*. Retrieved from http://eric.ed.gov/PDFS/ED399798.pdf
Kaderavek, J. N., Gillam, R. B., Ukrainetz, T. A., Justice, L. M., & Eisenberg, S. N. (2004). School-age children's self-assessment of oral narrative production. *Communication Disorders Quarterly, 26*, 37–48. doi:10.1177/15257401040260010401
Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: Praeger.
Kasanen, K., & Räty, H. (2002). 'You be sure now to be honest in your assessment': Teaching and learning self-assessment. *Social Psychology of Education, 5*, 313–328. doi:10.1023/A:1020993427849
Kasanen, K., Räty, H., & Eklund, A.-L. (2009). Elementary school pupils' evaluations of the malleability of their academic abilities. *Educational Research, 51*, 27–38.
Kirby, N., & Downs, C. (2007). Self-assessment and the disadvantaged student: Potential for encouraging self-regulated learning? *Assessment and Evaluation in Higher Education, 32*, 475–494.
Kuncel, N. R., Credé, M., & Thomas, L. L. (2005). The validity of self-reported grade point averages, class ranks, and test scores: A meta-analysis and review of the literature. *Review of Educational Research, 75*, 63–82.
Kwok, D. C., & Lai, D. W. (1993, May). *The self-perception of competence by Canadian and Chinese children*. Paper presented at the annual convention of the Canadian Psychological Association, Montreal, QC.
Laveault, D., & Miles, C. (2002, April). *The study of individual differences in the utility and validity of rubrics in the learning of writing ability*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

LaVoie, J. C., & Hodapp, A. F. (1987). Children's subjective ratings of their performance on a standardized achievement test. *Journal of School Psychology, 25*, 73–80. doi:10.1016/0022-4405%2887%2990062-8

Leahy, S., Lyon, C., Thompson, M., & Wiliam, D. (2005). Classroom assessment minute by minute, day by day. *Educational Leadership, 63*(3), 18–24.

Lew, M., Alwis, W., & Schmidt, G. (2010). Accuracy of students' self-assessment and their beliefs about its utility. *Assessment & Evaluation in Higher Education, 35*, 135–156.

Lipnevich, A., & Smith, J. (2008). *Response to assessment feedback: The effects of grades, praise, and source of information* (Research Report RR-08-30). Princeton, NJ: Educational Testing Service.

Lopez, R., & Kossack, S. (2007). Effects of recurring use of self-assessment in university courses. *International Journal of Learning, 14*, 203–216.

Luyten, H., & Dolkar, D. (2010). School-based assessments in high-stakes examinations in Bhutan: A question of trust? Exploring inconsistencies between external exam scores, school-based assessments, detailed teacher ratings, and student self-ratings. *Educational Research and Evaluation, 16*, 421–435.

Maki, R. H., Shields, M., Wheeler, A. E., & Zacchilli, T. L. (2005). Individual differences in absolute and relative metacomprehension accuracy. *Journal of Educational Psychology, 97*, 723–731.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: Macmillan.

Miller, T. L., Duffy, S. E., & Zane, T. (1993). Improving the accuracy of self-corrected mathematics homework. *The Journal of Educational Research, 86*, 184–189.

Miller, T. M., & Geraci, L. (2011). Training metacognition in the classroom: The influence of incentives and feedback on exam predictions. *Metacognition and Learning, 6*, 303–314. doi:10.1007/s11409-011-9083-7

Ng, J. R., & Earl, J. K. (2008). Accuracy in self-assessment: The role of ability, feedback, self-efficacy and goal orientation. *Australian Journal of Career Development, 17*, 39–50.

Nietfeld, J. L., Cao, L., & Osborne, J. W. (2005). Metacognitive monitoring accuracy and student performance in the postsecondary classroom. *The Journal of Experimental Education, 74*, 7–28.

Paulhus, D. L. (1991). Measurement and control of response bias. In J. Robinson, P. Shaver, & L. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (Vol. 1, pp. 17–59). New York, NY: Academic Press.

Raider-Roth, M. B. (2005). Trusting what you know: Negotiating the relational context of classroom life. *Teachers College Record, 107*, 587–628.

Ramdass, D., & Zimmerman, B. J. (2008). Effects of self-correction strategy training on middle school students' self-efficacy, self-evaluation, and mathematics division learning. *Journal of Advanced Academics, 20*, 18–41.

Ross, J. A. (2006). The reliability, validity, and utility of self-assessment. *Practical Assessment Research & Evaluation*, *11*(10), 1–13. Retrieved from http://pareonline.net/getvn.asp?v=11&n=10

Ross, J. A., Hogaboam-Gray, A., & Rolheiser, C. (2002). Student self-evaluation in grade 5–6 mathematics effects on problem-solving achievement. *Educational Assessment, 8*, 43–58.

Ross, J. A., Rolheiser, C., & Hogaboam-Gray, A. (1998a, April). *Impact of self-evaluation training on mathematics achievement in a cooperative learning environment*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Ross, J. A., Rolheiser, C., & Hogaboam-Gray, A. (1998b). Skills training versus action research in-service: Impact on student attitudes to self-evaluation. *Teaching and Teacher Education, 14*, 463–477.

Ross, J. A., Rolheiser, C., & Hogaboam-Gray, A. (2002). Influences on student cognitions about evaluation. *Assessment in Education: Principles, Policy & Practice, 9*, 81–95.

Saavedra, R., & Kwun, S. K. (1993). Peer evaluation in self-managing work groups. *Journal of Applied Psychology, 78*, 450–462. doi:10.1037/0021-9010.78.3.450

Sadler, P. M., & Good, E. (2006). The impact of self- and peer-grading on student learning. *Educational Assessment, 11*, 1–31.

Schunk, D. H., & Pajares, F. (2004). Self-efficacy in education revisited: Empirical and applied evidence. In D. M. McInerney & S. van Etten (Eds.), *Big theories revisited* (pp. 115–138). Greenwich, CT: Information Age.

Stipek, D. J. (1981). Children's perceptions of their own and their classmates' ability. *Journal of Educational Psychology, 73*, 404–410. doi:10.1037/0022-0663.73.3.404

Stipek, D. J., & Tannatt, L. M. (1984). Children's judgments of their own and their peers' academic competence. *Journal of Educational Psychology, 76*, 75–84. doi:10.1037/0022-0663.76.1.75

Sung, Y.-T., Chang, K.-E., Chang, T.-H., & Yu, W.-C. (2010). How many heads are better than one? The reliability and validity of teenagers' self- and peer assessments. *Journal of Adolescence, 33*, 135–145. doi:10.1016/j.adolescence.2009.04.004

Tierney, R. (2010). Fairness in classroom assessment. In H. Andrade & G. Cizek (Eds.), *Handbook of formative assessment* (pp. 125–144). New York, NY: Routledge.

Wall, S. M. (1982). Effects of systematic self-monitoring and self-reinforcement in children's management of test performances. *The Journal of Psychology, 111*, 129–136.

Wilson, J., & Wright, C. R. (1993). The predictive validity of student self-evaluations, teachers' assessments, and grades for performance on the verbal reasoning and numerical ability scales of the differential aptitude test for a sample of secondary school students attending rural Appalachia schools. *Educational and Psychological Measurement, 53*, 259–270. doi:10.1177/0013164493053001029