Adaptive Comparative Judgment in Graphics Applications and Education

Dr. Scott R. Bartholomew, Purdue University

I have instructed classes related to all CTE areas at the Junior High, High School, and College Level over the past 10 years. In addition to research activities I enjoying working with future and current Engineering/Technology Teachers. My interests revolve around adaptive comparative judgment, engineering design, teacher training, self-directed learning, and mobile devices in K-12 classrooms.

Dr. Patrick E. Connolly, Purdue University, West Lafayette (College of Engineering)

Dr. Patrick Connolly is a professor and department head of the Department of Computer Graphics Technology in the Purdue Polytechnic at Purdue University. He has extensive experience in the aerospace design and CAD/CAE software industries, and has been serving in higher education for twenty years. Dr. Connolly has a BS degree in Design and Graphics Technology and an MS in Computer Integrated Manufacturing from Brigham Young University, and a PhD in Educational Technology from Purdue University. His research interests include spatial ability development, virtual and augmented reality applications, product data and lifecycle management, and innovative classroom methodologies.

Adaptive Comparative Judging in Graphics Applications and Education

Abstract

One of the fundamental advantages behind Adaptive Comparative Judging (ACJ) is that it is easier and more accurate to comparison judge a series of products, and to develop a rank order of achievement, than it is to score products using a more subjective method or rubric approach. Research in the field of comparative judging has shown very high levels of reliability and close correlations between traditional grading approaches and this assessment methodology. This assessment approach appears to be effective at varying levels of rigor and academic achievement. Studies have examined adaptive comparative judging techniques in academic areas such as writing/composition, science education, and geography instruction. The areas of design and technology have proven to be especially effective topics for ACJ assessment, and are of special interest to the authors.

This introductory paper examines the fundamental principles of comparative judging and adaptive comparative judging, and discusses some of the most recent and relevant research on this topic. Key web-based ACJ tools and products are briefly reviewed—especially as they relate to academic settings. Applications in the areas of portfolio evaluation, graphics assessment, and peer critiquing are also explored.

Adaptive comparative judging has proven to be a method or assessment tool that is relatively straightforward to learn for faculty, and somewhat easily applied to a wide variety of topics and assignment approaches. ACJ appears to have a promising future in design and graphics applications.

The Problem with Open-ended Problems

Open-ended problems, a hallmark of many academic areas, are commonly employed in classrooms as a means of eliciting creativity (Kimbell, 2007), challenging students (Katehi, Pearson, & Feder, 2009), and fostering interest and engagement (Neal, 2011). The ability to work in and with ill-structured scenarios is a highly-sought-after skill among today's employers (Partnership, 2011; Resnick, Monroy-Hernandez, Rusk, Eastmond, Brennan, … & Kafai, 2009). However, open-ended problems carry with them the burden of a very difficult assessment task for the teacher—the inherently large number of possible solutions, the myriad of steps for completion, and the space for creativity in each assignment has proven very difficult for teachers to assess with validity and reliability (Kimbell, 2007; Pollitt, 2004). More potential for creativity by students can lead to a "messier" assessment scenario for teachers as the potential for a wide-range of solutions increases (Pollitt & Crisp, 2004).

Movements towards rubrics, criterion-based approaches, and technology-enhanced methodologies have all been lauded as potential options to alleviate some of the difficulties inherent in open-ended problem solving (Kimbell, 2007, 2012a; Denson, Buelin, Lammi, & D'Amico, 2015; Schilling & Applegate, 2012). Despite these approaches there is still no consensus as to best approaches for assessing open-ended problems with validity, reliability, and

efficiency (Pollitt, 2012a). Teacher/grader bias, subjectivity, and the many possible solutions inherent in open-ended problems make a purely rubric or criterion-based approach difficult to implement with fidelity—especially when more than one teacher/grader is involved in the assessment process (Bartholomew, 2017; Pollitt, 2004).

Adaptive Comparative Judgment

In 1927 Louis L. Thurstone published a paper on the law of comparative judgment. Thurstone (1927) argued that while humans have great difficulty making quality judgments with validity and reliability we are much more adept at making comparative judgments—judgments of quality between two items. In the early 2000s two researchers in the United Kingdom, Alastair Pollitt and Richard Kimbell, began to leverage Thurstone's ideas of comparative judgment in an effort to alleviate some of the difficulty related to open-ended design problems (Kimbell, 2012a). Rather than using a rubric or criterion-guide to tally points for students' assignments, a teacher would act as a judge and simply make a comparative judgment between two pieces of student work—essentially viewing two items and choosing the better of the two based on their own expertise and a predetermined rubric or criteria. As a teacher repeats this process—making comparative judgments between different pairs of items—each item of student work begins to rise or fall, in the overall rank-order, based on their performance. Over time pieces of student work gain a "win-loss" record; each time an item is chosen over another the piece of student work it counts as a "win," while a "loss" stems from not being chosen when paired with another item (Pollitt, 2004, 2012a; 2012b).

Recent advances in technology have facilitated the creation of multiple adaptive comparative judgment (ACJ) software applications and platforms. These systems are useful in facilitating and streamlining the ACJ process as teachers can simply view two items on a computer screen and choose the better of the two. Technology advancements have opened the door for multiple types of student work (i.e., images, documents, audio/video recordings) to be compared using ACJ. Starting with Kimbell and Pollitt's work and moving forward ACJ has been piloted, tested, and refined over time and the algorithm which facilitates the judgments has been improved (Pollitt 2004, 2012b). Comparative judgment becomes "adaptive" as the algorithm and technology tools work towards pairings becoming increasingly refined—rather than random pairing items with similar "win-loss" records are compared with one another and the overall rank order is increasingly refined in terms of validity and reliability (Pollitt, 2004).

Validity

As with any assessment the issue of validity, or how well an instrument measures what it was intended to measure, must be addressed (Gall, Borg, & Gall, 1996). Approaches to investigating the validity of ACJ have revolved around identifying the correlation between the emerging rank-order from ACJ and scores obtained through traditional grading approaches. Obviously these approaches work on the assumption that our current grading approaches and practices are valid. In the United Kingdom the ACJ ranking from various settings has been compared with the results of traditional grading methods (Kimbell, Wheeler, Miller, & Pollitt, 2007) and has shown high-levels of consistency between assessment approach ($R^2 = 0.81$, corresponding with a correlation of 0.90). Similarly, in the United States Bartholomew, Strimel, & Jackson (2017)

compared the resulting ACJ rank order and the original rubric-based grading scores from several student projects and found a high degree of correlation ($r_s = .80$, $p < .01$). Similarly, Seery, Buckley, Doyle, & Canty (2016) identified high correlations between the grading criteria scores—obtained through traditional marking—and the final rank-order obtained through ACJ for graphic capability. Acting under the assumption that traditional grading approaches are valid, with the understanding that a full discussion of assessment validity is beyond the scope of this paper, these findings suggest that ACJ is similarly valid as it has produced highly-correlated results to traditional marking. Further research into the validity of ACJ as an assessment approach would shed additional light on its' feasibility for widespread implementation.

Current ACJ technologies also prompt judges/teachers to provide a comment following each judgment which justifies their decision to choose one item over another. These comments help capture the rationale and decision-making process of the judges and can be used as a validity-check when compared with the assignment rubric. Studies have demonstrated a close alignment between the assignment rubric and the judge rationale for judgment (Strimel et al., 2017) further strengthening the validity of ACJ as a tool for assessment. In addition to strengthening the validity of the final rank order the judge comments "follow" each item and can be provided to students at the conclusion of grading as a form of feedback regarding their work and its' place in the final rank order.

Reliability

Pollit (2004, 2012b) pointed out that one of the biggest challenges with traditional methods of assessment is reliability or inconsistency in results when multiple graders or grading sessions are employed (Gall, Borg, & Gall, 1996). This unreliability further intensifies in scenarios that involve open-ended problems which incorporate an aspect of design—personal preferences, the order in which items are graded, and a variety of other factors contribute to the overall difficulty in grading with reliability in these settings (Pollitt, 2004). Pollitt (2004) noted that certain subject areas, those areas emphasizing design, creativity, and open-ended problems, struggle with reliability more than other areas: "problems like this [with reliability in assessment] seem to occur most prominently in certain less traditional subject areas such as Information and Communications Technology and aspects of Design and Technology" (p. 5).

The literature related to the reliability of ACJ as a method of assessment has consistently identified higher levels of reliability ($r > .7$) than those reached through other methods of assessment (Pollitt, 2004). Noting that micro-judgments, rubrics, criteria-guides, and multiple graders have all been used in an effort to increase assessment reliability, Pollitt (2004) points to the unique strength of ACJ for reliable assessment:

> When a judge compares two performances (using their own personal 'standard' or internalized criteria) **the judge's standard cancels out**. In theory the same relative judgment is expected from any well-behaved judge. A similar effect occurs in sport: when two contestants or teams meet the 'better' team is likely to win, whatever the absolute standard of competition and irrespective of the expectations of any judge who might be involved. The result of the comparisons of this kind is objective relative measurement. (pp. 6-7, emphasis in original)

Many of the technology tools related to ACJ also have inherent abilities to check for inter-rater reliability as "rogue" judges and items can be "flagged" for further training and review (Pollitt, 2004); if a judge consistently grades in a manner contradictory to other judges that judge can be flagged for training. Similarly, if an item of student work is not ranked consistently (i.e. if the item is ranked near the top of the items but does not "win" in a comparison with items near the bottom) it can be flagged for further comparison and review. These functionalities, built into current ACJ technology products can help strengthen the inter-rater reliability of the overall results (Pollitt, 2004).

During the ACJ process a misfit statistic is computed for each judge and each item. The *Rasch* model misfit statistic for each judge represents how closely the decisions for that particular judge align with the final rank order of student work. The *Rasch* model misfit statistic for each item represents how consistently each item was ranked in comparison with other items (see Pollitt, 2012b for a discussion of the *Rasch* statistical methods in ACJ). The misfit statistics can be used as a reliability check for judges and the final rank order of items with further analysis and assessment of problematic judges or items.

In addition to the measures covered, this method of assessment has demonstrated strong stochastic transitivity (if A usually beats B, and B usually beats C, then A will mostly beat C), furthering strengthening the reliability of the findings (Pollitt, 2004). Pollitt (2004) pointed out that the strong reliability findings connected with ACJ account for possible unreliability between graders as well as lack of internal consistency within the assignment itself—an uncommon characteristic as most traditional reliability coefficients only allow for one of these.

ACJ in Practice

The majority of work to date with ACJ has revolved around design and technology settings (Kimbell, 2012b; Pollitt, 2004). However, recent efforts have emphasized the use of ACJ for assessment in engineering design (Bartholomew, Reeve et al., 2017; Strimel et al., 2017), graphic capability (Seery, et al., 2016), design portfolios (Bartholomew, Reeve et al., 2017; Hartell & Skogh, 2015; Strimel, et al., 2017), teacher training and assessment (McMahon & Jones, 2015), and other academic areas including geography, chemistry, biology, accounting, psychology, sociology, English, Math, health, social care, business, speaking, and foreign language (Pollitt, 2004, 2012a). In addition to academic settings ACJ has found a home in industry as a method for comparing various iterations of products as well as employee training and competency-based evaluation ("DigitalAssess," 2017). The body of research consistently demonstrates that open-ended scenarios, which provide an opportunity for the effective implementation of ACJ, are not confined to one particular subject area or genre of work (Pollitt, 2012a), rather, new approaches and arenas for implementation of ACJ are, and should continue to be, piloted, studied, and evaluated.

To a great extent, the application of ACJ in the design/graphics/media realm is a natural progression of the long-standing efforts around portfolio-based learning and assessment. Portfolios (and more recently, e-portfolios) have been an effective tool for the collection and display of artifacts (Lorenzo & Ittelson, 2005). They are commonly used in learning settings,

professional development, and other scenarios in both formative and summative assessment (Klenowski , Askew, & Carnell, 2006). Barrett (2005) noted that there are multiple definitions and uses for portfolios, including learning, assessment, employment, sales and marketing, and collection of individuals' best work. As learning/teaching tools, portfolios can be applied in both achievement of standards and constructivist/iterative processes (Avraamidou & Zembal-Saul, 2002; Barrett, 2005; Klenowski, et al., 2006).

However, it is well-established that portfolio assessment can be a complex and time-consuming procedure when attempted with traditional, rubric-based methods. ACJ provides a strong, reliable, and valid alternative to subjective and rubric-based approaches (Tarricone & Newhouse, 2016). In the product design process field, the dynamic, iterative, and free-flowing nature of learning require an assessment technique that provides a clear and unambiguous view of holistic accomplishment (Seery, Buckley, Doyle, & Canty, 2016; Seery, Canty, & Phelan, 2012). In applying these same principles in a more inclusive sense to a broader view of 'graphics' applications, it seems logical that ACJ could be an effective assessment tool for digital portfolios in the realms of media graphics, data visualization, video applications, and digital art. Each of these areas, and related others within the computer graphics umbrella, struggle with the challenge of accurate assessment of student artifacts.

An additional advantage of the application of ACJ in graphics fields is the positive results that emerge from the use of peer evaluation in student learning. Evaluation by peers provides opportunities for higher-order cognitive understanding of learning outcomes and applications, and provides students an opportunity to reflect on their individual efforts (Jones & Alcock, 2014; Seery & Delahunty, 2013). Comparative judging approaches provide the ability to assess the broader and more divergent results that often occur in graphics-related artifacts, thereby enriching the deep learning that peer-to-peer review encourages (Jones & Wheadon, 2015).

The majority of research around ACJ has been centered in Europe (Hartell & Skogh, 2015; Kimbell, 2012a, 2012b; Seery, Canty, & Phelan, 2012) with recent efforts in the United States (Bartholomew, Reeve et al., 2017; Strimel, et al., 2017) and Australia (Heldsinger & Humphry, 2010). While ACJ has been utilized in K-12 classrooms the majority of ACJ-use has taken place with higher education, awarding bodies, and industry (M. Wingfield, personal communication, November 15, 2016). Collaborative efforts to use ACJ in international settings have recently been undertaken with future efforts planned for international assessment research.

Tools for ACJ

Currently there are limited options for ACJ assessment using free and paid-for products commercially developed and marketed to companies and individuals. Each of the products has slight variations but encompass similar capabilities, interfaces, and outputs. *CompareAssess*, owned and marketed by *DigitalAssess* was the first to commercially market and capitalize on ACJ as a tool for assessment following the research efforts of Kimbell and Pollitt (digitalassess.com). Use of the *CompareAssess* ACJ engine requires a fee for licenses and use and is currently largely centered on business and industry applications.

No More Marking, another similar company offers free limited use of their ACJ engine with further capabilities available for purchase (nomoremarking.com). *No More Marking* originated from the Digital Platform for the Assessment of Competences (D-PAC), a consortium dedicated to providing an open-source comparative judgment application. *No More Marking* appears to be more accessible for educators and individuals looking to complete ACJ on a smaller scale while *CompareAssess* has a broader range of possible outputs for larger-scale projects. The code from the D-PAC group is published online and freely available at Github (https://github.com/ubc/compare) and a demonstration of the code from D-PAC can be found at: https://demo.d-pac.be/.

Final Thoughts

ACJ would likely not be feasible for classroom implementation without the technology tools to expedite and automate the comparative judgment process. Despite the challenges in implementing ACJ without currently available technologies it is not impossible—several teachers already employ comparative judgment techniques without realizing it (Bartholomew, 2017). Some teachers place student work in "piles" according to relative quality while others assign an initial grade to one assignment and then modify that grade based on the other assignments—both of these are forms of comparative judgment already employed by teachers.

While anecdotal evidence from several ACJ studies suggests that using ACJ for assessment *may not significantly* alter the amount of time required for the grading process of open-ended assignments (Bartholomew, Reeve et al., 2017; Strimel et al., 2017), other research has demonstrated markedly increased time requirements for ACJ (Tarricone & Newhouse, 2016). What ACJ does accomplish revolves around increasing the reliability of the final results and rank-order of student work (Pollitt, 2004, 2012b). The inherent biases from teacher perceptions, experiences, expectations, and differences can all be addressed through the ACJ process (Bartholomew, 2017; Pollitt, 2004).

Several studies related to ACJ have analyzed the time taken by each judge to complete each judgment. Interestingly these studies have consistently shown that taking more time to complete judgments does not correlate with better judgments - as demonstrated by how well a judges decisions align with the final rank order achieved (Bartholomew, Reeve et al., 2017; Kimbell, 2017; Strimel et al., 2017). As judges rely on the provided rubric and their own professional expertise their ability to make comparative judgments more efficiently may increase (Kimbell, 2017) and the process of assessment through ACJ may improve into a more efficient method of grading – especially in open-ended scenarios.

It seems probable that ACJ will see increased relevance in the assessment of broader media graphics portfolios based on its strengths in design product review, open-ended problem environments, and peer critiquing. As Rowsome, Seery, Lane, and Gordon (2013) stated, "Adaptive Comparative Judgment is a dynamic assessment tool to facilitate and capture the complex iterative design process" (p. 23.1185.2). Additionally, ACJ appears to encourage students' creativity and enhance higher-order learning, both of which are necessary for development in the graphics/design/digital media domains. These attributes make ACJ a promising, holistic assessment tool for this complex realm. The authors' intent going forward is

to apply ACJ methods to the assessment of digital portfolios in the areas of computer game design and development, animation, data visualization, video compositing and effects, computer graphic simulation, and digital art. To our knowledge, ACJ has had limited exposure in these broader fields of media graphics. To this end, a preliminary research project will be implemented that will apply ACJ in an introductory computer graphics course during the fall 2017 academic semester. The targeted course provides a foundation for the development and use of raster and vector images for a variety of applications. Full-color images, illustrations, and basic animations are produced using computer technologies, with a focus on both technical and aesthetic aspects. Topics addressed in the course include color theory and perception, surface and lighting analysis, rendering techniques, and other technical characteristics of effective graphic production. It is believed that these components will provide a robust and interesting test for ACJ methodologies. On-going research will look at ACJ applications with increasing complex design artifacts produced in higher level graphics courses in the fields mentioned here.

References

Avraamidou, L., & Zembal-Saul,C. (2002). Making the case for the use of web-based portfolios in support of learning to teach. *The Journal of Interactive Online Learning*, *1* (2) 1-19.

Barrett, H. C. (2005). Researching electronic portfolios and learner engagement. The Reflect Initiative; Researching Electronic Portfolios: Learning, Engagement, Collaboration, through Technology. Retrieved from http://ww.w.electronicportfolios.org/reflect/whitepaper.pdf

Bartholomew, S. R. (2017). Assessing open-ended design problems, *Technology and Engineering Education Teacher*, *76*(6), pp. 13-17

Bartholomew, S. R., Reeve, E., Veon, R., Goodridge, W., Stewardson, G., Lee, V., Nadelson, L. (2017). Mobile devices, self-directed learning, and achievement in Technology and Engineering Education classrooms during a STEM activity. *Journal of Technology Education* (accepted for publication).

Bartholomew, S. R., Strimel, G. S., & Jackson, A. (2017). *A comparison of traditional and adaptive comparative judgment assessment techniques for freshmen engineering design projects.* Manuscript submitted for publication.

Denson, C. D., Buelin, J. K., Lammi, M. D., & D'Amico, S. (2015). Developing instrumentation for assessing creativity in engineering design. *Journal of Technology Education, 27*(1), pp. 23-40

DigitalAssess, 2017. *Vocational learning*. Retrieved from http://digitalassess.com/vocational-learning/

Gall, M. D., Borg, W. R., & Gall, J. P. (1996). *Educational research* (6th ed.). White Plains, NY: Longman.

Hartell, E., & Skogh, I. B. (2015). Criteria for success: A study of primary technology teachers' assessment of digital portfolios. *Australasian Journal of Technology Education, 2*(1).

Heldsinger, S., & Humphry, S. (2010). Using the method of pairwise comparison to obtain reliable teacher assessments. *The Australian Educational Researcher, 37*(2), 1-19.

Jones, I., & Alcock, L. J. (2014). Peer assessment without assessment criteria. *Studies in Higher Education, 39* (10), 1774-1787. doi:10.1080/03075079.3013.821974

Jones, I., & Wheadon, C. (2015). Peer assessment using comparative and absolute judgement. *Studies in Educational Evaluation, 47*, 93-101.

Katehi, L., Pearson, G., & Feder, M. (2009). Engineering in K-12 education. Committee on K-12 Engineering Education, National Academy of Engineering and National Research Council of the National Academies.

Kimbell, R. (2007). E-assessment in project e-scape. *Design & Technology Education: An International Journal, 12*(2), 66-76.

Kimbell, R. (2012a). The origins and underpinning principles of e-scape. *International Journal of Technology & Design Education*, 22, 123-134.

Kimbell, R. (2012b). Evolving project e-scape for national assessment. *International Journal of Technology & Design Education*, 22, 135-155.

Kimbell, R. (2017). *Making assessment judgements: policy, practice, and research*. Manuscript in preparation

Kimbell, R., Wheeler, T., Miller, S., & Pollitt, A. (2007). *E-scape portfolio assessment phase 2 report*. London, England: TERU, Goldsmiths, University of London.

Klenowski, V., Askew, S., & Carnell, E. (2006) Portfolios for learning, assessment and professional development in higher education. *Assessment & Evaluation in Higher Education, 31*(3), 267-286. doi: 10.1080/02602930500352816

Lorenzo, G., & Ittelson, J. (2005). *An overview of e-portfolios*. Educause Learning Initiative. Retrieved from https://net.educause.edu/ir/library/pdf/eli3001.pdf

McMahon, S., & Jones, I. (2015). A comparative judgement approach to teacher assessment. *Assessment in Education: Principles, Policy & Practice, 22*(3), 368-389.

Neal, M. A. (2011). Engaging students through effective questions. *Education Canada, 51*(1), n1.

Partnership. (2011). *Framework for 21st century learning.* Retrieved from http://www.p21.org/our-work/p21-framework

Pollitt, A. (2004). Let's stop marking exams. Retrieved from http://www.cambridgeassessment.org.uk/images/109719-let-s-stop-marking-exams.pdf

Pollitt, A. (2012a). Comparative judgement for assessment. *International Journal of Technology and Design Education, 22*(2), 157-170

Pollitt, A. (2012b). The method of adaptive comparative judgement. *Assessment in Education: Principles, Policy & Practice, 19*(3), 281-300.

Pollitt, A., & Crisp, V. (2004). *Could comparative judgments of script quality replace traditional marking and improve the validity of exam questions?* Paper presented at the BERA Annual Conference, UMIST Manchester, England.

Resnick, M., Maloney, J., Monroy-Hernández, A., Rusk, N., Eastmond, E., Brennan, K., ... & Kafai, Y. (2009). Scratch: Programming for all. *Communications of the ACM*, *52*(11), 60-67.

Rowsome, P., Seery, N., Lane, D., & Gordon, S. (2013). The development of pre-service design educator's capacity to make professional judgments on design capability using Adaptive Comparative Judgment. *Proceedings of the 120th ASEE Annual Conference & Exposition* (23.1185.1-23.1185.10). Atlanta: American Society for Engineering Education.

Schilling, K., & Applegate, R. (2012). Best methods for evaluating educational impact: a comparison of the efficacy of commonly used measures of library instruction. *Journal of the Medical Library Association: 100*(4), 258–269. http://doi.org/10.3163/1536-5050.100.4.007

Seery, N., Buckley, J., Doyle, A., & Canty, D. (2016). The validity and reliability of Adaptive Comparative Judgements in the assessment of graphical capability. *Proceedings of the 71st Mid Year Conference of the Engineering Design Graphics Division* (104-109). Nashua, NH: American Society for Engineering Education.

Seery, N., Canty, D., & Phelan, P. (2012). The validity and value of peer assessment using adaptive comparative judgement in design driven practical education. *International Journal of Technology and Design Education*, *22*(2), 205-226.

Seery, N., & Delahunty, T. (2013). Capturing graphical capability through Ipsative enquiry using Adaptive Comparative Judgement. *Proceedings of the 68th Mid Year Conference of the Engineering Design Graphics Division* (48-52). Worcester, MA: American Society for Engineering Education.

Strimel, G., Bartholomew, S. R., Jackson, A., Grubbs, M., Bates, D., & Kim, E. (2017). *Evaluating freshman engineering design projects using Adaptive Comparative Judgment.* Manuscript in preparation

Tarricone, P., & Newhouse, C. P. (2016). Using comparative judgement and online technologies in the assessment and measurement of creative performance and capability. *International Journal of Educational Technology in Higher Education 13*(16). doi: 10.1186/s41239-016-0018-x

Thurstone, L. L. (1927). A law of comparative judgment, *Psychological Review*, *34*, 273-286