Источники снижения надёжности оценки заданий с развёрнутым ответом по иностранным языкам

Бодоньи Марина Алексеевна

кандидат педагогических наук, доцент кафедры английского языка в профессиональной сфере ФГБОУ ВО «Кубанский государственный университет», г. Краснодар
m.bodony@mail.ru

Ключевые слова: оценка письменной речи, оценка продуктивных видов речи, задания с развёрнутым ответом, надёжность оценки, качество оценки.

Оценка продуктивных видов иноязычной речевой деятельности характеризуется сложной организацией, что обусловлено различными компонентами, составляющими структуру ситуации оценки, и влиянием многочисленных факторов на процесс оценивания. Центральной фигурой данного процесса является эксперт, выступающий в качестве субъекта оценки, от точного и объективного решения которого зависит результат квалиметрического анализа иноязычного высказывания. «Феномен оценки проявляется не только как процесс (оценивание) и его результат (оценка), но и как свойство, способность, психическая функция субъекта деятельности..., как форма интерпретации и конструирования»[1]. Изучение деятельности экспертов охватывает широкий круг проблем, к которым относятся определение качества и точности оценки, выявление причин снижения надёжности, коррекция случаев нарушения оценочной деятельности и смещения оценки, уточнение стратегий достижения точности, валидности и объективности квалиметрического анализа высказываний при учёте человеческого фактора и т.п. В данном контексте представляется целесообразным привести мнение Дж. Коннор-Линтона, который подчёркивает необходимость изучения деятельности эксперта: «Если мы не знаем, что делают эксперты..., тогда мы не можем знать, что означают их оценки»[2].

В предлагаемой статье представлен обзор исследований, посвящённых изучению типичных случаев снижения надёжности оценки, имеющих место в ходе квалиметрической интерпретации иноязычного письменного высказывания, на основе чего осуществляется попытка их систематизации и анализа. Снижение надёжности оценки соотносится с феноменом смещения оценки, рассматриваемым нами как нарушение оценочной деятельности, приводящее к неточным результатам оценивания. Учитывая систематический характер проявления данной тенденции[3], очевидна необходимость особого внимания к факторам, влияющим на недостаточную объективность и надёжность результатов оценки,

[1] *Сутужко В.В.* Феномен оценки в социальном бытии и познании: автореф. дис. … докт. филос. наук / В.В. Сутужко: 09.00.11. — Саратов: СГУ, 2006. — 36 с. — С. 8.

[2] *Connor-Linton, J.* Looking behind the curtain: What do L2 com position ratings really mean? / J. Connor-Linton // TESOL Quarterly. — 1995. — No 29. — P. 762–765. — p.763.

[3] *Wigglesworth, G.* Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction / G. Wigglesworth // Language Testing, 1993. — No 10(3). — P. 305–335. — p. 309.

в том числе, касающихся причин вариативного поведения эксперта.

Достижение качественных параметров оценки рассматривается в контексте соблюдения «принципов инвариантного измерения»[4], основанных на независимости оценки и минимизации нежелательного влияния отдельных факторов. Однако квалиметрический анализ продуктов письменной речи относится к так называемой оценке, опосредованной субъектом контроля. Таким образом, оцениваемые письменные высказывания воспринимаются и интерпретируются через призму его опыта, представлений, ожиданий и т.п. Представляется очевидной необходимость учёта человеческого фактора, потенциально снижающего надёжность оценки. Концептуализация оценки, опосредованной субъектом контроля, предполагает наличие двух фундаментальных основ: с одной стороны, необходима разработка теоретических основ квалиметрической процедуры, ориентированной на уточнение шкал, критериев и т.п., с другой стороны, особую важность приобретает исследование субъективных процессов, сопровождающих оценочную деятельность[5]. Разработка этих двух теоретических направлений способствует повышению надёжности оценки продуктов письменной речи, минимизации смещения оценки как следствие изучения источников нарушений и разработки приёмов их предупреждения и коррекции.

Исследователи выделяют различные источники снижения надёжности оценки заданий с развёрнутым ответом: факторы, порождаемые продуктом письменной речи, экспертом, оценивающим продукты письменной речи, и ситуацией оценки[6]; ошибки, обусловленные оценочным инструментарием, индивидуальными характеристиками эксперта, случайными факторами[7].

Б. О'Салливан разграничивает физические/физиологические, психологические и профессиональные характеристики эксперта как основание для выявления вариативности его поведения и источник снижения межэкспертной надёжности. К физическим/физиологическим характеристикам относятся краткосрочные (зубная боль, холод и т.п.) и долгосрочные (проблемы со зрением, слухом и т.п.) недомогания, а также возраст, пол; к психологическим — память, когнитивный стиль, особенности эмоционального поведения, мотивация и т.п.; профессиональные параметры связываются с уровнем образования, специальной подготовкой к оценочной деятельности, уровнем владения языком и опытом преподавательской и оценочной деятельности[8].

К. Мерфи и Дж. Клевелэнд разграничивают три группы критериев, используемых для анализа качественных параметров оценочной деятельности, и, соответственно, указывающих на возможные причины снижения надёжности оценки: традиционные психометрические критерии; ошибки эксперта как следствие субъективности, предвзятости и т.п., в результате чего наблюдается смещение оценки; непосредственные ошибки точности оценивания, связанные с используемым квалиметрическим инструментарием[9].

Анализ и обобщение теоретических источников показывает, что к наиболее существенным источникам снижения надёжности оценки иноязычной письменной речи относятся оценочный инструментарий, процедура оценки, эксперт, оценивающий иноязычное письменное высказывание.

Нарушения, связанные с оценочным инструментарием, могут быть вызваны неэргономичностью оценочных шкал вследствие, например, несоблюдения границ дробности

---

[4] *Engelhard, G., Stefanie, Jr.A.* Wind Invariant measurement with raters and rating scales: Rasch Models for Rater-Mediated Assessments / G. Engelhard, Jr. A. Stefanie. — New York: Routledge, 2018. — 368 p.

[5] *Engelhard G., Stefanie Jr.A.* Wind Invariant measurement with raters and rating scales: Rasch Models for Rater-Mediated Assessments / G. Engelhard, Jr. A. Stefanie. — New York: Routledge, 2018. — 368 p.

[6] *Weigle S.C.* Effects of training on raters of ESL compositions / S. C. Weigle // Language Testing. — 1994. — Vol. 11. — Issue 2. — P. 197–223.

[7] *Bachman L.* Fundamental Considerations in Language Testing / L. Bachman. — Oxford: Oxford University Press, 1990. — 420 p.

[8] *O'Sullivan, B.* Towards a model of performance in oral language tests/ B. O'Sullivan: Unpublished PhD Dissertation. — CALS: University of Reading, 2000. — 324 p.

[9] *Murphy, K., Cleveland, J.* Performance appraisal: An organizational Perspective / K. Murphy, J. Cleveland. — Boston: Allyn & Bacon, 1991. — 349 p. — p. 211–212.

оценочных шкал, нарушения предела градаций, неясных и расплывчатых дескрипторов и т.п.[10]

Процедура квалиметрической интерпретации иноязычного письменного высказывания может стать источником смещения оценки при условии нарушения её системной организации и недостаточного развития функциональных компонентов ситуации оценки. Необходимо подчеркнуть, что деятельность эксперта имеет нелинейный и рекурсивный характер[11], вместе с тем, это не может расцениваться как отсутствие системности. Моделирование деятельности эксперта, оценивающего письменные работы, позволяет дифференцировать стратегии и этапы оценивания. Так, А. Камминг, Р. Кантор и Д. Е. Пауэрс определяют оценивание как деятельность, ориентированную на принятие решений. Она включает стратегии интерпретации, обеспечивающие формирование представления об оцениваемой работе, и стратегии оценки, предполагающие формулирование оценочных суждений. В каждой группе стратегий дифференцируется направленность внимания эксперта на разные уровни деятельности:

- уровень самоконтроля;
- уровень интерпретации и оценки риторических и содержательных аспектов;
- уровень анализ языковых характеристик интерпретируемого текста[12].

Соответственно снижение надёжности оценки заданий с развёрнутым ответом может иметь место на каждом уровне, что связано с несоответствием применяемых стратегий. На уровне самоконтроля ошибки эксперта связаны с невнимательным чтением задания и письменного высказывания, недостаточной концентрацией на критериях, тенденцией сравнения работ между собой, неполным разграничением аспектов совмещения критериев, ограниченным обоснованием оценочного решения и т.п.

На риторическом и содержательном уровне нарушение оценочной деятельности происходит вследствие ложной интерпретации неясных или двойственных в смысловом отношении фраз, неточного распознавания структуры и организации текста, чрезмерно широкого или узкого обобщения идей, некорректной оценки логики рассуждения, предвзятой оценки степени выполнения задания и развития темы и т.п.

На уровне интерпретации языковых средств эксперт может иметь затруднения при разграничении степени «грубости» языковых ошибок, их частотности, определении правильности использования лексических, грамматических, синтаксических и т.п. средств, корректности орфографии и пунктуации, что также ведёт к снижению надёжности оценки.

Кроме того, внешние характеристики процедуры оценки могут отрицательно влиять на поведение эксперта, что создаёт условия для искажения результатов оценочной деятельности. Например, использование значительного количества объектов контроля в процедуре оценки приводит к их размыванию и нечёткой дифференциации, следствием чего становится неточность оценки. Значительная нагрузка влияет на утомление, что в свою очередь оказывает отрицательное влияние на когнитивные процессы эксперта, его внимание, память и т.п., что снижает надёжность оценки. Также использование компьютерных технологий, обеспечивающих проверку продуктов письменной речи субъектами оценки в режиме онлайн, может быть проблематичным как в плане отсутствия профессионального взаимодействия субъектов оценки между собой вследствие удалённой проверки, так и в плане собственно адаптации к требованиям онлайновых систем.

Таким образом, оценочная процедура имеет как внешние, так и внутренние аспекты, обусловливающие потенциальную возможность снижения надёжности оценки продуктов иноязычной письменной речи. Представляется целесообразным их учёт на основе рациональной органи-

---

[10] *Яновский М.И.* Проблема оптимального объёма оценочной шкалы как инструмента деятельности педагога / М.И. Яновский // Психологическая наука и образование. — 2013. — № 3. — С. 18–25.

[11] *Baker, B.* An Updated Visual Representation for Writing Assessment Research/ B. Baker // Canada Journal TESL. — 2016. — Volume 32. — P. 124–136.

[12] *Cumming, A.* Scoring TOEFL Essays and TOEFL 2000 Prototype Writing Tasks: An Investigation into Raters' Decision Making and Development of a Preliminary Analytic Framework/ A. Cumming, R. Kantor, D.E. Powers // Educational Testing Service. — Princeton, New Jersey, 2001. — 201 p.

зации условий оценивания, уточнения используемых инструментов и средств контроля. Что касается эффективного использования экспертом стратегий оценочной деятельности как компонента процедуры оценивания, по нашему мнению, необходимо развитие внутренней согласованности эксперта, которая рассматривается не только как «тенденция оценивать одинаково одно и тоже письменное высказывание в разных случаях»[13], но и как способности в заданном интервале времени в конкретных условиях в установленных пределах к выполнению функций по поддержанию заданного режима работы. Так, по мнению Дж. М. Линакра, межэкспертная согласованность оценок не может рассматриваться как единственная цель, обеспечивающая надёжность оценивания, в то время как важным представляется обучение экспертов для принятия самосогласованных решений[14]. Одним из направлений развития внутренней надёжности эксперта является дифференциация приёмов оценки продуктов письменной речи, понимание рациональности использования тех или иных стратегий на отдельных этапах деятельности и самоанализ эффективности оценочного процесса и точности его результатов.

Эксперт, оценивающий продукты иноязычной письменной речи, выступает как носитель «атрибутов личности»[15], детерминирующих поведенческие проявления в процессе интерпретации письменного текста, что в определённой степени оказывает влияние на качество оценки. К подобным атрибутам исследователи относят статус языка эксперта по отношению к оцениваемым продуктам речи (родной или иностранный), пол, уровень образования и специальную подготовку к осуществлению оценочной деятельности, опыт оценивания письменных высказываний и т.п.

Рассмотрение поведения эксперта как источника снижения надёжности оценки продуктов письменной речи предполагает исследование феномена субъективности как потенциальной черты, характеризующей оценку, опосредованную человеком.

Субъективность, проявляющаяся в процессе оценивания, имеет двойственный характер. С одной стороны, письменное высказывание как творческий и индивидуальный продукт претерпевает индивидуальную интерпретацию со стороны эксперта. Даже при наличии критериев оценки и детализированных дескриптивных шкал имеет место субъективное восприятие, влияющее на результат оценки. С другой стороны, сами критерии оценки могут интерпретироваться индивидуально каждым экспертом, что также является следствием проявления субъективности. В данном контексте представляется целесообразным привести цитату Р. Гамароффа о том, что «одинаковые баллы, выставленные экспертами, не обязательно означают идентичные оценочные суждения»[16].

Кроме того, субъективность, проявляющаяся в неспособности к точным оценкам, беспристрастным и непредвзятым выводам, соотносится с моделями поведения эксперта в процессе квалиметрической интерпретации письменного высказывания. В рамках проводимого нами исследования представляется очевидной необходимость их дифференциации.

Эффект снисходительности/строгости предполагает, что эксперт ставит более высокие или наоборот более низкие оценки, чем того заслуживает продукт письменной речи. Таким образом, в первом случае оценки группируются в верхней части шкалы, во втором — в нижней. Г. Энгельхард рассматривает данную поведенческую модель как континуум, т.е. постепенно меняющуюся последовательность от снисходительности к строгости[17], что, с одной стороны, репрезентирует данный эффект как сравнительную величину, характеризующую

[13] *Weigle, S.C.* Assessing writing/ S. C. Weigle. — Cambridge: Cambridge University Press, 2002. — 278 p. — p. 135.

[14] *Linacre, J.M.* Many-faceted Rasch measurement / J.M. Linacre. — Chicago, IL: MESA Press, 1989. — 158 p.

[15] *Bachman L.F., Palmer A.S.* Language assessment in practice: Developing language assessments and justifying their place in the real world / L.F. Bachman, A.S. Palmer. — Oxford, UK: Oxford University Press, 2010. — 510 p.

[16] *Gamaroff, R.* Teacher reliability in language assessment: The bug of all bears / R. Gamaroff // System. — 2000. — No 28. — P. 31–53.

[17] *Englehard, G.* Examining rater errors in the assessment of written compositions with a many-faceted Rasch model / G. Englehard // Journal of Educational Measurement. — 1994. — No 31(2). — P. 93–112.

поведение эксперта, с другой стороны, свидетельствует об изменчивости его поведения. Кроме того, эксперт может интерпретировать разные аспекты письменного высказывания с разной степенью строгости, например, быть более строгим в оценке языковых средств и снисходительным к содержательным аспектам. Выявление эффекта снисходительности/ строгости связано с использованием статистических методов, к которым относятся сравнение средних оценок со средними точками используемых рейтинговых шкал; использование метода дисперсионного анализа (ANOVA) для определения того, существует ли статистически значимый эффект и изучение степени асимметрии частотных распределений оценок[18].

Эффект ореола проявляется как тенденция к генерализации, представляющая склонность оценивать продукты письменной речи на основе общего впечатления, причинами чего, как правило, являются недостаточно сформированные умения дифференцировать степень репрезентации аспектов письменного высказывания в соответствии с критериями оценки и предлагаемой шкалой и склонность устанавливать один и тот же уровень для всех аспектов. Таким образом, эксперт обобщённо рассматривает «концептуально отличные и независимые аспекты письменного высказывания и выставляет сходные баллы»[19]. Исследователи указывают на то, что эффект ореола может проявляться как объективная характеристики и может иметь иллюзорный характер. В первом случае эффект ореола возникает как реальное относительное совмещение оцениваемых аспектов, во втором случае он является следствием недостатков оценочных средств, наблюдений, используемых стратегий и т.п. Соответственно разграничение двух типов эффекта ореола предполагает различные подходы к их учёту в оценочной деятельности. Безусловно, второй тип требует особого внимания для его предупреждения и возможной коррекции оценочного поведения эксперта. Представляется интересным, что эффект ореола не всегда ведёт к нарушению объективности оценки, а напротив, как показывает исследование В.Х. Купера, может способствовать активизации точности оценки. Данное явление получило название «парадокс точности эффекта ореола»[20]. Как можно заметить, хотя эффект ореола и не обязательно подразумевает низкий уровень точности, необходимость внимания к данному феномену в процессе оценивания продуктов письменной речи очевидна для повышения надёжности и объективности оценки.

Эффект усреднения, проявляющийся как особенность поведения эксперта в процессе квалиметрической интерпретации письменного высказывания, соотносится с тенденцией выбора средней категории рейтинговой шкалы при избегании крайних точек. Данная тенденция определяет ограничения ранжирования обучающихся, следствием чего становится снижение надёжности и достоверности оценки, что соответственно ведёт к смещению результатов оценочной деятельности. Данный эффект, как правило, обусловлен недостаточной внутренней согласованностью субъекта оценки, его сомнениями и неуверенностью при интерпретации письменных высказываний. Причинами проявления рассматриваемого эффекта могут быть недостаточная дифференциация экспертом критериальной шкалы, игнорирование максимальных баллов в целях стимулирования обучающихся к достижению более высокого уровня, нежелание выставлять высокие баллы работам, которые оцениваются первыми (при рассмотрении большого количества письменных высказываний), избегание крайне низких баллов как боязнь ошибки и т.п.

Эффект последовательности основан на тенденции субъекта оценки сравнивать оцениваемые продукты иноязычной письменной речи между собой: в таком случае оценки, выставленные за предыдущие работы, оказывают влияние на квалиметрическую интерпретацию последующих.

[18] *Saal, F.E.* Rating the ratings: Assessing the psychometric quality of rating data / F.E. Saal, R.G. Downey, M.A. Lahey // Psychological Bulletin. — 1980. — No 88. — P. 413–428.

[19] *Myford, M., Wolfe, E.W.* Detecting and Measuring Rater Effects Using Many-Facet Rasch Measurement: Part II / M. Myford, E. W. Wolfe // Journal of Applied Measurement. — 2003. — No 5(2). — P. 189–227.

[20] *Cooper, WH.* Ubiquitous halo / WH. Cooper // Psychological Bulletin. — ГОД. — No 90. — P. 218–244.

Эффект последовательности имеет два типа репрезентации: эффект контраста, когда оценка слабых письменных высказываний может служить основой для более высокой оценки последующих работ и наоборот, и эффект ассимиляции, когда под влиянием высоких оценок слабые работы также получают высокие рейтинги и наоборот[21]. Рассматриваемые особенности поведения эксперта ведут к снижению надёжности оценки, обусловленному сравнением работ между собой и принятием неточных решений.

Таким образом, представленные модели поведения эксперта в процессе квалиметрической интерпретации иноязычного письменного высказывания обусловливают снижение надёжности оценки и нарушение точности её результатов. Систематизация случаев смещения оценки иноязычной письменной речи, основанная на выделении потенциальных источников искажений (оценочная шкала, процедура оценки, субъект оценки), позволяет уточнить теоретическое описание оценочного процесса и специфику деятельности субъекта оценки.

На основе анализа теоретических источников нами была предпринята попытка систематизировать факторы снижения надёжности оценки иноязычной письменной речи. К ним относятся:

- процедурные;
- инструментальные;
- операционные;
- поведенческие (на уровне самооценки);
- поведенческие (на уровне оценки содержания);
- поведенческие (на уровне оценки языковых средств);
- когнитивные;
- физиологические;
- эмоциональные;
- атрибутивные склонности.

Нами были выделены две крупные группы источников, обуславливающих смещение оценки: случайные и системные. Случайные определяются нами как независимые переменные, которые сложно предугадать, они — заранее неизвестны и зависят от ряда случайных обстоятельств. В качестве примеров случайных источников можно привести сбои технического плана и форс-мажорные обстоятельства. Представляется очевидным, что в определённых условиях они могут оказывать существенное влияние на снижение надёжности, но в силу того, что проявление случайных факторов непредсказуемо, мы не конкретизируем их типы: в зависимости от конкретного случая они могут варьироваться.

Системные факторы подразделяются на две крупные группы: внешние, связанные с компонентами ситуации оценки, и внутренние, определяемые фигурой и оценочной деятельностью эксперта. Внешние источники охватывают процедурные и инструментальные аспекты. Процедурные аспекты детерминированы последовательностью действий, механизмами, процедурами, алгоритмами и т.п., используемыми для оценивания письменных высказываний, также они связаны с нагрузкой и собственно условиями организации процесса оценивания. Инструментальные, в свою очередь, соотносятся с оценочными средствами, например, критериями, шкалами, подходами и т.п. В качестве примера влияния оценочного инструментария на снижение надёжности можно привести неясные формулировки или частичное совпадение критериев, что ведёт к затруднениям оценивания параметров текста. Кроме того, к инструментальным аспектам может быть отнесён тип оценки (холистический, аналитический и т.п.), оказывающий влияние на возможное смещение оценки.

Внутренние источники снижения надёжности охватывают процессные и личностные аспекты. Процессные, проявляющиеся непосредственно в ходе оценивания, делятся на операционные и поведенческие. Операционные включают технические, технологические и тактические источники смещения оценки. Технические связаны с описками, арифметическими ошибками, некорректными подсчётами, неправильными записями и т.п. Технологические проявляются как несоответствие приёмов оценивания методическим рекомендациям, следование неизменной технологии несмотря на разные форматы/ условия порождения письменного

[21] *Attali, Y.* Sequential effects in essay ratings/ Y. Attali. — Educational and Psychological Measurement. — 2011. — No 71. — Pp. 68–79.

высказывания и т.п. Тактические источники связаны с отсутствием постоянного контроля за степенью межэкспертной надёжности и эффективностью оценивания, с переоценкой внутренней надёжности эксперта, кроме того они проявляются в виде нарушений тактики оценивания, например, когда эксперт концентрирует внимание на одном аспекте текста. Поведенческие источники снижения надёжности оценки проявляются в тесной связи со стратегиями, используемыми экспертами в процессе оценивания письменного высказывания, и реализуются на уровне самоконтроля, на уровне интерпретации содержания и анализа языковых средств.

Личностные характеристики рассматриваются нами как индивидуальные переменные, обуславливающие смещение оценки иноязычного письменного высказывания. Это — когнитивные, физиологические, фоновые, эмоциональные характеристики и атрибутивные склонности.

Предложенная нами типологии стала основой для разработки модели анализа сниженной надёжности оценки заданий с развёрнутым ответов по иностранным языкам. Опора на элементы, представленные выше, целесообразна в условиях нарушения межэкспертной согласованности для определения источников снижения надёжности оценки иноязычного письменного высказывания.

Мы выделили их основные группы и конкретизировали случаи, обусловленные спецификой появления снижения надёжности. Обратим внимание на то, что предложенные нами случаи не могут рассматриваться как окончательный список: для составления их полного перечня требуется проведение дальнейших исследований.

В заключение следует отметить, что разработанная нами типология источников снижения надёжности оценки иноязычного письменного высказывания может выступать теоретико-эмпирической основой для проведения научных исследований в области контроля продуктивных видов иноязычной речи и накопления знаний о ситуациях, условиях и факторах смещения точности и объективности оценки. Безусловно, основная цель разработки типологии — её применение для предупреждения и профилактики возникающих случаев нарушения оценочной деятельности и разработки эффективных способов их коррекции.