ГЕНЕРАЛИЗАЦИЯ ЗНАЧЕНИЙ ВЫБОРОЧНЫХ КОЭФФИЦИЕНТОВ ВАЛИДНОСТИ ТЕСТОВЫХ РЕЗУЛЬТАТОВ[1]

Хои Суен,
Пенсильванский университет США
HoiSuen@psu.edu
Джуху Ким,
Аджойский университет,
Южная Корея
juhu@ajou.ac.kr

Валидность тестовых результатов (Validity) определяется как мера соответствия интерпретированных данных тестирования предполагаемым целям, эмпирическим данным, а также содержанию теории, на основе которой создаётся тест. Это определение дано согласно профессиональным стандартам, установленным Объединенным Комитетом по стандартам педагогического и психологического тестирования[2].

Валидность результатов одного и того же теста может меняться от исследования к исследованию и в зависимости от различных выборочных совокупностей испытуемых. Помимо этого, на изменения значений коэффициентов валидности могут влиять и другие различные факторы. Объединенный Комитет определил пять главных источников обоснования валидности тестовых результатов. Это:

1) мера соответствия содержания теста целям тестирования;

2) такая организация процесса тестирования, которая не допускает искажения результатов;

3) адекватность цели тестирования структуры самого теста;

4) мера соответствия реальных последствий тестирования ранее выдвинутым целям;

5) определение меры связи результатов тестировании с результатами по другим показателям (переменным величинам).

1 Статья написана специально для Российского журнала «Педагогические Измерения». Сокращённый перевод с английского языка Светланы Янченко.

2 Этот Комитет объединяет американскую педагогическую и психологическую ассоциации исследователей, а также Национальный совет по педагогическим измерениям (Американскую Образовательную Ассоциацию Исследователей (1999).

Как бы ни менялась концепция валидности за прошедшие 80 лет, один из перечисленных источников валидизации тестовых результатов применялся неизменно. Это определение меры связи результатов теста с другими показателями. Тем самым делались попытки определить — как результаты теста коррелируют с некоторыми внешними критериями успешности. Типичный пример такого обоснования валидности тестовых результатов — это коррелирование тестовых балов, получаемых при приёме в вуз, с показателями учебных достижений в процессе дальнейшего в нём обучения. Чем выше корреляция, тем выше уверенность, что результаты абитуриентских тестов помогают повысить качество студентов. Другой пример — корреляция между оценками дошкольников и оценками младших школьников. В таких случаях корреляция используется для оценки валидности дошкольных оценок для выявления, в последующем, детей с определёнными учебными затруднениями. И третий пример — вычисление коэффициента корреляции между баллами по тесту интеллектуальности (IQ) с тестом для оценки способностей рассуждать, аргументировать и обосновывать (то, что по-английски называется Verbal reasoning test). Это делается для того, чтобы показать, что интеллект действительно включает в себя упомянутые способности.

Процесс обоснования валидности тестовых результатов начинается с подбора подходящей выборки испытуемых и получения результатов тестирования в этой совокупности. Затем начинается поиск данных по критерию, в качестве которых используются реальные результаты. Например, для абитуриентского теста критерием могут быть результаты текущих учебных достижений студентов, сдававших абитуриентский тест. Или, другой пример, оценки учебных достижений школьников могут использоваться для валидизации тестовых результатов оценки способностей у дошкольников. При этом обычно считается т.н. классический коэффициент корреляции Пирсона. Получаемый в таких случаях коэффициент корреляции иногда называют коэффициентом валидности тестовых результатов. Высокое значение такого коэффициента корреляции, если это сочетается и с другими формами обоснования, даёт неоспоримое свидетельство валидности результатов проверяемого теста.

Различные коэффициенты валидности результатов одного и того же теста

Одно из затруднений относительно статистического обосно-

вания валидности тестовых результатов состоит в том, что невозможно точно определить ту выборку испытуемых, для которых разрабатывается тест[3]. Общее правило, которым руководствуются при валидизации результатов — это организовать тестирование репрезентативной выборки из генеральной совокупности тех испытуемых, для которых предназначен разрабатываемый тест. Но практически сделать это невозможно, а потому у каждого разработчика может оказаться своя выборка испытуемых, которую можно назвать локальной. Естественно возникает вопрос получения обобщённой оценки коэффициента валидности результатов теста.

Например, разработчику абитуриентского теста в качестве испытуемых могут рассматриваться все потенциальные выпускники школ. Однако для разработки качественного теста для приёма в конкретный колледж интерес представляет только те испытуемые, которые поступают именно в данный колледж, и более того, только те, кто поступает на конкретную специальность этого учебного заведения. Таким образом, состав выборок испытуемых и условия сбора данных могут меняться у разных разработчиков теста. Следствием этого получаются различающиеся коэффициенты валидности тестовых результатов. Коэффициенты валидности отличаются даже в одних и тех же выборках, что было подтверждено в работах Ghiselli, 1966; Lent, Aurbach, and Levin, 1971. Когда различия в коэффициентах становятся заметными, возникает вопрос интерпретации. Как доверять локальному коэффициенту валидности? Проблемная ситуация с различиями в значениях валидности тестовых результатов рассматривалась в работах Linn, Harnisch, & Dunbar, 1981; Peralman, Schmidt, & Hunter, 1980. До 1990-х, доминировала точка зрения, что в каждая выборка имеет свои особенности, а потому и значения коэффициенты валидности тестовых результатов имею смысл и значение только для данных выборок, и вряд ли могут обобщены (генерализованы). Schmidt, 1988, рекомендовал периодически отслеживать изменения выборочных коэффициентов валидности.

Однако, в литературе есть альтернативное объяснение тому, почему различные выборки испытуемых порождают различные значения коэффициентов валидности тестовых результатов. Было выдвинуто предположение, что эти значения зависят не столько от уникальных различий выборок испытуемых, сколько от обычной вариации выборочных статистик и случайных ошибок измерений.

[3] Иногда такую группу называют целевой (target group). (Прим. переводчика).

Schmidt and Hunter (1977) предложили метод генерализации значений выборочных коэффициентов валидности тестовых результатов, что является, по сути, единственным, в своём роде, методом применением мета-анализа для поиска ответа на вопрос какое истолкование зависимости значений коэффициентов валидности правильнее брать за основу при определении истинного значения коэффициента валидности результатов данного теста: являются ли различия коэффициентов валидности следствием различий условий формирования локальных выборочных совокупностей, или это обычные статистические артефакты. Если верно первое, то в центр внимания надо ставить локальные выборочные коэффициенты валидности, если же верно второе, то это означает возможность выведения обобщенного значения коэффициента валидности, а значит, нет и необходимости проводить множество отдельных выборочных исследований.

Общий план генерализации локальных выборочных значений коэффициентов валидности тестовых результатов

Schmidt and Hunter предложили оценить возможные источники статистических артефактов в отдельных выборочных исследованиях, что позволить их элиминировать и, тем самым, получить обобщенную оценку коэффициента валидности (Hunter, Schmidt, & Jackson, 1982; Hunter & Schmidt, 1990). Если после удаления статистических артефактов различия коэффициентов валидности в выборочных исследованиях заметно уменьшаются, то это предполагает, что есть одно общее значение коэффициента валидности результатов теста, проявляющее себя в выборочных исследованиях. Если же после устранения статистических артефактов вариация в значениях выборочных коэффициентов всё ещё остаётся существенной, то это означает, что различия в локальных выборках являются главным источником изменения коэффициентов валидности тестовых результатов, а потому коэффициенты валидности надо будет по-прежнему считать в отдельных выборках, и не более.

Если говорить терминами модели дисперсионного анализа, то общая вариация наблюдаемых локальных выборочных коэффициентов валидности тестовых результатов зависит от:

1) различий в коэффициентах надежности тестовых результатов, что влияет на изменение ошибочной части вариации тестовых результатов (test reliability);

2) различий в надежности критериев, которые тоже могут влиять на возникновения ошибочных компонентов оценки коэффициентов валидности тестовых результатов (criterion reliability);

3) различий в размахе (range) тестовых результатов в каждой выборке, что также влияет на значения дисперсии баллов.

4) различий в дисперсии, вызванных ошибками выборки (due to random sampling errors);

5) различий формирования каждой локальной выборочной совокупности испытуемых.

Из этих пяти источников вариации значений выборочных коэффициентов валидности тестовых результатов первые четыре являются статистическими артефактами, которые можно учесть и устранить. И если после этого вариация выборочных коэффициентов становится незначительной, то можно определённо вывести, что причина отмеченной вариации — указанные статистические артефакты. А это означает возможность наличия одного общего значения коэффициента валидности для всех выборочных исследований. И что локальные выборки — это выборки, результаты в которых можно использовать для получения практически полезной генерализованной оценки валидности тестовых результатов.

Если появляется возможность получения одного общего генерализованного значения валидности тестовых результатов на основе локального выборочного исследования, то легко вывести ненужность множества выборочных исследований валидности тестовых результатов,. Если же, однако, вариация значений выборочных коэффициентов валидности тестовых результатов остаётся по прежнему большой, то это может означать, что всё дело в существенных различиях локальных выборок. А потому придется продолжать практику обоснования валидности тестовых результатов посредством множества сравнительных выборочных исследований для каждого теста[4].

Устранение артефактов, возникающих вследствие недостаточной надёжности тестовых результатов и различий в значениях размаха данных

Первые два артефакта могут быть элиминированы посредством применения формул коррекции на понижение вариации тестовых результатов, полученных в выборке (correction for attenuation). Третий артефакт можно устранить формулами коррекции на уменьшение размаха данных. Эти известные ме-

4 То , что в западной литературе называется cross-validization.

тоды объединены здесь в общую методику генерализации значения выборочного коэффициента валидности тестовых результатов.

Вначале делается коррекция на понижение выборочного коэффициента валидности тестовых результатов из-за недостаточной их надёжности.

Для этого используется формула

$$a_{1i} = \frac{1}{\sqrt{\rho_{11_i}}}, \qquad (1)$$

где ρ_{11_i} — это значение коэффициента надёжности тестовых результатов по тесту в локальной выборке *i*.

Затем, похожая коррекция делается для данных, используемых в качестве критерия (a_{2i}) для выборочного коэффициента валидности тестовых результатов.

$$a_{2i} = \frac{1}{\sqrt{\rho_{22_i}}}, \qquad (2)$$

где ρ_{22_i} — коэффициент надежности критерия, используемый в локальной выборке под номером *I*. Добавим к этому коррекцию (a_{3i}) на снижение размаха выборочных данных (по сравнению с размахом данных в генеральной совокупости), по формуле

$$a_{3i} = \frac{u_i}{\sqrt{1+(u_i^2-1)r_i^2}} \qquad (3)$$

где r_i^2 — это квадрат коэффициента валидности в локальной выборке, а u_i есть отношение оценки стандартного отклонения генеральной совокупности к оценке данных в локальной выборочной совокупности, под номером *I, получаемое из формулы*

$$u_i = \frac{\sigma_{unrestricted\ range}}{\sigma_{restricted\ range}}. \qquad (4)$$

Влияние этих артефактов, связанных с ненадёжностью теста и критерии, а также с необходимостью коррекции на снижения размаха данных, корректируют каждый локальный коэффициент валидности тестовых результатов посредством произведения

$$r_{i_C} = a_{1i}a_{2i}a_{3i}r_i, \qquad (5)$$

где становится корректированным коэффициентом валидности тестовых результатов в локальной выборке.

Другие методы коррекции выборочных оценок коэффициентов валидности тестовых результатов

Коррекция четвертого статистического артефакта несколько сложнее трёх предыдущих. Математически трудно оценить влияние фактора расчёта коэффициента валидности в локальной выборочной совокупности на значение коэффициента валидности в генеральной совокупности. Для минимизации влияния погрешности локаль-

ной выборочной совокупности на значение коэффициента валидности тестовых результатов в генеральной совокупности можно использовать так называемый эмпирический Байесовский процесс т.н. стабилизации значения коэффициента валидности тестовых результатов в локальной выборке. Эта стабилизация выполняется методом взвешенного усреднения оценки выборочного коэффициента валидности и предполагаемого значения этого коэффициента в генеральной совокупности. Значение предполагаемого генерализованного коэффициента валидности получают усреднением всех имеющихся в наличии выборочных оценок коэффициентов валидности тестовых результатов.

В качестве весов для получения взвешенной средней арифметической коэффициента валидности генеральной совокупности используются значения единицы, делённой на значение дисперсии ошибочного компонента вариации коэффициента валидности, полученного в соответствующей локальной выборке результатов[5]. Другими словами, чем больше стандартная ошибка коэффициента валидности, тем меньшим становится вклад этого локального коэффициента при расчёте коэффициента валидности в генеральной совокупности. Если локальные значения коэффициентов валидности получаются в малых в выборках испытуемых, то значения статистической ошибки этого коэффициента становятся большими. Если же выборки большие, то возрастает и их весомость.

В результате попытки стабилизировать выборочные коэффициенты валидности посредством использования взвешенных средних арифметических выборочных коэффициентов валидности и коэффициента валидности в генеральной совокупности, значения этих выборочных коэффициентов меняются. После чего, меняется, соответственно, и предполагаемое значение коэффициента валидности в генеральной совокупности. Эти изменения повлекут за собой пересчет взвешенных выборочных значений коэффициентов валидности тестовых результатов. По мере изменения значений коэффициентов процесс может повторяться, позволяя каждый раз улучшать точность оценок, до тех пор, пока различия между значениями в выборках и генеральной совокупности не совпадут. После чего процесс конвергенции оценок заканчивается. Получаемые при этом оценки выборочных коэффициентов валидности являются наиболее репрезентативными.

До начала описанного процесса конвергенции, выборочные значения коэффициен-

5

This is actually an implementation of Bayesian statistics for the normal process in which the global coefficient is treated as the prior value and the local coefficient as the sample statistic. The resulting weighted average is equivalent to a Bayesian posterior estimate, which is theoretically more accurate than either the prior or the sample statistic.

тов валидности необходимо перевести в z-преобразование Фишера, что необходимо сделать для минимизации влияния смещённых распределений выборочных коэффициентов валидности на последующие расчеты взвешенных средних значений.

$$z_i = \frac{\ln\left(\frac{1+r_{i_C}}{1-r_{i_C}}\right)}{2}, \qquad (6)$$

где r_{i_C} коэффициент валидности выборочных тестовых результатов преобразован к значению на шкале Фишера.

Значение дисперсии трансформированных таким образом коэффициентов валидности $\hat{\sigma}^2_{z_i}$, можно получить по формуле:

$$\hat{\sigma}^2_{z_i} = \frac{a^2_{1i} a^2_{2i} a^2_{3i}}{n_i - 3}, \qquad (7)$$

где n_i — число испытуемых в выборке под номером j, используемой для определения коэффициента валидности, полученного на данной выборке испытуемых.

Значение коэффициента в генеральной совокупности и значения эмпирического процесса Байеса оцениваются посредством z-преобразований Фишера. Итеративный процесс опирается на известный алгоритм E-M (Estimation and Maximization, Dempster, Laird, & Rubin, 1977). Этот алгоритм использует двухшаговый циклы. В каждом цикле определяемая оценка максимизируется. В цикле максимизации все z_i, используются для получения оценки коэффициента валидности в генеральной совокупности.

$$\hat{\mu}_z = \frac{\sum_{i=1}^{k} z_i}{k}. \qquad (8)$$

Значение дисперсии, ассоциируемое с распределением вероятных значений коэффициента валидности тестовых результатов в генеральной совокупности равна

$$\hat{\sigma}^2_{\mu_z} = \frac{\sum_{i=1}^{k} [z_i^2 + \hat{\sigma}^2_{z_i}]}{k} - k\hat{\mu}^2, \qquad (9)$$

где k — есть число выборочных исследований по определению коэффициента валидности тестовых результатов.

В шаге E делаются попытки улучшить оценку каждого значения z_i и $\hat{\sigma}^2_{z_i}$ посредством эмпирического метода Байеса, где используются оценки коэффициента валидности генеральной совокупности, полученные в шаге E предыдущего цикла. Статистически это означает возможность получения улучшенных оценок выборочных коэффициентов валидности использованием оценок предыдущего цикла:

$$\text{improved } z_i = \frac{z_i \hat{\sigma}^2_{\mu_z} + \hat{\mu}_z \hat{\sigma}^2_{z_i}}{\hat{\sigma}^2_{z_i} + \hat{\sigma}^2_{\mu_z}}. \qquad (10)$$

Дисперсия значений улучшенных выборочных коэффициентов валидности равна

$$\text{improved } \hat{\sigma}^2_{z_i} = \frac{1}{\frac{1}{\hat{\sigma}^2_{z_i}} + \frac{1}{\hat{\sigma}^2_{\mu_z}}}. \quad (11)$$

На шаге М каждой новой итерации улучшенные z_i и

$$\text{improved } \hat{\sigma}^2_{z_i}$$

используются для последующего уточнения, что приводит, в свою очередь, улучшению оценок z_i и $\hat{\sigma}^2_{z_i}$ посредством применения формул (10) и (11) в очередном Е-шаге. Окончательные оценки значений z_i становятся скорректированными выборочными коэффициентами валидности тестовых результатов, из которых устранены ошибки измерения, вызванные всеми пятью упоминавшимися факторами.

Обсуждение результатов

Поскольку влияние всех пяти упомянутых статистических артефактов было минимизировано, в окончательных оценках z_i, все остающиеся различия между выборочными коэффициентами валидности тестовых результатов можно объяснить фактором специфичности выборки (factor of unique local conditions), в которой был получен каждый коэффициент. Поскольку стандартные ошибки выборочных значений коэффициентов валидности тестовых результатов было минимизированы, но не элиминированы полностью, то даже если бы теоретически существовал бы только один коэффициент валидности для генеральной совокупности, окончательные значения выборочных z_i не были бы всё равно идентичными. Неизбежно оставалась бы некоторая вариация в окончательных оценках z_i значений. Если остаточная дисперсия не существенна, то мы вправе полагать, что такая вариация вызвана влиянием случайного отбора испытуемых в локальных выборках. Теория статистики в качестве генерального параметра допускает существовании только одного значения коэффициента корреляции, которое может быть сопоставлено со значениями выборочных коэффициентов. Для науки важна оценка генерального параметра, а потому знать множество локальных значений выборочных коэффициентов валидности не нужно. Если же остаточная вариация остаётся заметной, то это означает существенную значимость различий выборочных данных, и необходимость специального анализа причин такого положения.

Каково граничное значение для принятия решения о несущественности остаточной вариации? Hunter, Schmidt and Jackson (1982) предложили так называемое 75%-е решающее

правило, в соответствии с которым если 75 или более процентов исходной дисперсии локальных коэффициентов валидности объяснимы влиянием статистических артефактов, то можно принять гипотезу о существовании одного общего значения коэффициента валидности тестовых результатов в генеральной совокупности. Из чего следует, что нет необходимости проводить множество исследований валидности на локальных выборках. В этом выводе заключается заметная прагматическая значимость результатов настоящего исследования.

В противном случае придётся допустить, что коэффициенты валидности одного и того же теста, полученные в разных локальных выборках, разными исследователями, оказываются несопоставимыми в принципе. А значит и влияние испытуемых отдельных выборок на результаты теста оказываются статистически достоверными. Поэтому необходимо будет признать существенную научную значимость именно локальных выборочных коэффициентов валидности тестовых результатов.

Такой практически важный вывод можно сделать при сравнении дисперсии значений исходных коэффициентов валидности и оценкой $\hat{\sigma}^2_{\mu_z}$, полученной вследствие итерации по алгоритму E—M, до начала их конвергенции. Если дисперсия значений коэффициентов валидности в локальных выборках в четыре раза больше оценки $\hat{\sigma}^2_{\mu_z}$, то можно согласиться с идеей одного общего коэффициента валидности. Относительно применения этого довольно условного решающего правила ещё предстоит придти к консенсусу (James, Demaree, and Mulaik, 1986). Альтернативный подход — это проверить статистическую достоверность гипотезы отличия $\hat{\sigma}^2_{\mu_z}$ от нуля. Если нулевая гипотеза сохраняется, мы вправе сделать вывод, что остаточная дисперсия несущественна. Если же нулевая гипотеза отклоняется, то можно вывести, что существует необъяснимая часть дисперсии.

Другие аспекты

Следование 75% правилу или какому-либо другому критерию в случаях, когда остаточная часть дисперсии признаётся несущественной, даёт обоснование действовать исходя из идеи наличия одного общего коэффициента валидности тестовых результатов. Что делать, если эта часть остаётся большой? Означает ли это, что нам придётся каждый раз обосновывать валидность тестовых результатов в каждом случае? Что является причиной различий локальных значений коэффици-

ентов валидности? Если в случаях большой остаточной части дисперсии идея существования одного общего коэффициента валидности кажется неприемлемой, то быть может, можно допустить идею менее универсальной общности, типа существования общего коэффициента валидности для концептуально меньших генеральных совокупностей? Подход Schmidt и Hunter не даёт ответа на такого рода дополнительные вопросы.

Однако решение возможно. Процесс генерализации вывода о валидности тестовых результатов можно провести, опираясь на оценки z_i и $\hat{\sigma}^2_{z_i}$ (см. формулу 7). Последующий процесс минимизации значения стандартной ошибки коэффициентов валидности посредством Байесовского подхода можно сделать более эффективным, если к оценкам z_i and $\hat{\sigma}^2_{z_i}$ применить известный метод иерархического линейного моделирования (Hierarchical Linear Modeling, HLM, см. Bryk and Raudenbush, 1992). Одно из преимуществ применения HLM для проведения заключительного этапа процессса валидизации — это наличие статистической программы (HLM6 software, Raudenbush, Bryk, & Congdon, 2005), использующей Байесовский процесс. Это преимущество становится заметным, если значение остаточной дисперсии остаётся большой, статистически достоверной.

Процесс генерализации коэффициентов валидности тестовых результатов, включая метод HLM, представляет важную форму специального применения методов мета-анализа данных. (cf. Cooper, 1998; Glass, McGaw, & Smith, 1981; Hedges & Olkin, 1985; Rosenthal, 1991). Естественным образом, на этот процесс распространяются ограничения, присущие методам мета-анализа данных.

References

1. American Educational Research Association. Standards for Educational and Psychological Testing. Washington, DC: Author. 1999.
2. *Bryk A.S., Raudenbush S.* Hierarchical linear models: Applications and data analysis methods. Newbury Park: Sage. 1992.
3. *Raudenbush S., Bryk, A.S. & Congdon R.T.* HLM-6. Lincolnwood, IL: Scientific Software International. 2005.
4. *Cooper H.* Synthesizing research: A guide for literature reviews. Thousand Oaks, CA: Sage. 1998.
5. *Dempster A.P., Laird N.M., & Rubin D.B.* Maximum likeli-

20

Абрамова М.Е., Белобородов В.Н., Татур А.О. Оценка точности измерения способностей экзаменуемых. С. 142–151. Сб. «Оценка образовательных достижений в рамках национальных экзаменов». Материалы и тезисы докладов Международной конференции. 13–15 декабря 2004 г. М.: Изд-во «Уникум-центр», 2005. 279 с.

21

Там же.

22

Лемуткина М. Единый государственный обман. http://www.gazeta.ru/education/2006/01/30_a_528596.shtml 23.03.06.

23

Федеральная служба по надзору в образовании и науке. Стратегические цели и тактические задачи Федеральной службы по надзору в сфере образования и науки. http://www.obrnadzor.gov.ru/strategy 23.03.06.

hood estimation from incomplete data via the EM algorithm. Journal of Royal Statistical Society, Series B, 1977; 39: 1–38.
6. *Ghiselli E.E.* The validity of occupational aptitude tests. New York: Wiley. 1966.
7. *Glass G.V.* Meta-analysis at 25. Retrieved on May 20, 2006 from http://glass.ed.asu.edu/gene/papers/meta25.html. 2000.
8. *Glass G.V., McGaw B., & Smith M.L.* Meta-analysis in social research. Beverly Hills, CA: Sage. 1981.
9. *Hedges L.V., & Olkin I.* Statistical methods for meta-analysis. New York: Academic Press. 1985.
10. *Hunter J.E., Schmidt F.L., & Jackson G.B.* Meta-analysis: Cumulating research findings across studies. Beverly Hills, CA: Sage Publications. 1982.
11. *Hunter J.E., & Schmidt F.L.* Methods of meta-analysis: Correcting for sources of error and bias in research findings. Newbury Park, CA: Sage. 1990.
12. *James L.R., Demaree R.G., & Mulaik S.A.* A note on validity generalization procedures. Journal of Applied Psychology, 1986; 71: 440–450.
13. *Lent R.H., Aurbach H.A., & Levin L.S.* Predictors, criteria, and significant results. Personnel Psychology, 1971; 24: 519–533.
14. *Linn R.L., Harnisch D.L., & Dunbar S.B.* Validity generalization and situational specificity: An analysis of the prediction of first-grade year grades in law school. Applied Psychological Measurement, 1981; 5(3): 281–289.
15. *Pearlman, Schmidt F.L., & Hunter J.E.* Validity generalization results for tests used to predict job proficiency and training success in clerical occupations. Journal of Applied Psychology, 1980; 65(4): 373–406.
16. *Rosenthal R.* Meta-analytic procedures for social research (Rev. ed.). Thousand Oaks, CA: Sage. 1991.
17. *Schmidt F.L.* Statistical innovations in validity assessment. In Howard Wainer & Henry Braun (Eds.). *Test Validity*. New Jersey: Erlbaum. 1988; pp. 171–189.
18. *Schmidt F.L., & Hunter J.E.* Development of a general solution to the problem of validity generalization. Journal of Applied Psychology, 1977; 31: 215–231.