IJET

Chang, F, T. Y., & Li. M. N. F. (2019). A relative effectiveness assessment of an immediate feedback assessment technique employed on computers and scratch cards: a meta-analysis. *International Journal of Educational Technology, 6*(1), 12-29.

Original Paper

Journal homepage: http://educationaltechnology.net/ijet/

OPEN ACCESS

A relative effectiveness assessment of an immediate feedback assessment technique employed on computers and scratch cards: a meta-analysis

Frederic Tao-Yi Chang, Mao-Neng Fred Li*

National Chia-Yi University, Taiwan

KEYWORDS

Immediate feedback assessment technique
IF-AT
Meta-analysis
Multiple- choice testing

ABSTRACT

IF-AT (also called answer until correct) is a revised form of traditional multiple-choice testing. It was created to provide instant feedback during testing and to tap into partial knowledge of learners. The purposes of this meta-analytic study were to compare the effectiveness of IF-AT with the traditional multiple-choice test; to assess effectiveness of the computer-based IF-AT with the scratch card based IF-AT; and to identify modifies of IFAT effects. A total of 35 IFAT studies were examined. Data were analyzed using traditional meta-analysis and FAT-PET-PEESE-MRA. The results indicated: 1) The overall effect size of IF-AT over the traditional multiple choice was .581(fixed) and .618(random). By FAT-PET-PEESE MRA, the net overall effect size of IF-AT was .522, which was close to a previous meta-analysis result (Bangert-Drowns et.al., 1991) based on 4 answer-until-correct studies; 2) In terms of IF-AT forms, effectiveness of IFAT interacted with testing materials. Scratch card-based IF-AT was significantly more effective than computer-based IF-AT in language arts and social sciences (LS); while they worked equally well in mathematics and science (MS); and 3) Test materials was identified as a significant moderator to explain a large part of heterogeneity between IF-AT studies. In terms of test materials, students taking IF-AT tests always performed better in LS than MS regardless whether it was scratch card-based IF-AT or computer-based IFAT. The evidence suggests that IF-AT is very effective in enhancing students' learning within LS by memory retention, and it is less effective in MS of which requires more conceptual understanding than memory retention.

Background

The multiple-choice test is one of the most widely-used forms of assessment in education due to clear rubrics for assessment and easy administration in a large classroom, but it is often unable to measure students’ learning outcomes fairly. There are three shortcomings with the use of multiple-choice items (Epstein, Epstein, and Brosvic, 2001) in a formative test. Firstly, it cannot provide students with immediate feedback. Secondly, it is unable to assess

* *Corresponding author.* E-mail address: fredli@mail.ncyu.edu.tw

students' partial knowledge. Thirdly, it often fails to enhance students' learning during testing since students cannot verify the right answer without immediate feedback. As a result, their misconceptions may be retained, or even replace the correct concepts (Shute, 2007). What makes it even worse, a traditional multiple-choice test may discourage low-achievers in a test, due to its inability to measure students' partial knowledge (Beever et al., 1999).

As an improved assessment technique of multiple-choice tests, the scratch-card IF-AT (immediate feedback assessment technique) was introduced to fulfill what traditional multiple-choice tests could not offer. With its ability to measure students' partial knowledge, to obtain feedback on an item by item basis, and to increase their motivation, the scratch-card IF-AT could facilitate students' active involvement in learning (Epstein Enterprises, 2017). Some researchers did argue for its effectiveness, but the scratch-card IF-AT did not prevail in school settings before 1990 due to its inconvenience to develop and produce. Later, Bangert-Drowns et al. (1991) synthesized four empirical studies of carbon coating-based or latent image-based IF-AT by a meta-analysis. The overall magnitude of its effectiveness was .53 (Cohen's d). At the same time, Bangert-Drowns et al. (1991) reported the effect size of feedback aided by computers was .22(Cohen's d).

Besides, this meta-analysis did not include any studies conducted in either a scratch-card form or a computer-based form. Recently, due to the proliferating use of modern technology in daily life, IF-AT can be easily administered by computers. This alternative IF-AT format has been applied frequently in many universities, especially for both individual and group readiness assurance testing (Robinson, Sweet, & Mayrath, 2008). For example, Hattie (2009) estimated the effect size of computer-based intervening feedback instruction to be .52 (Cohen's d) based on 161 studies. Similarly, studies of effectiveness and advantages between the scratch-card IF-AT versus the computer-based IF-AT have often been addressed (Lopez, 2009). Although more studies have implemented feedback aided by computers, there has been no meta-analytic synthesis regarding the effect size of the computer-based IF-AT.

Moreover, there were contradictory results of relative effectiveness in the literature (Epstein, et al, 2002; Smiley, 2011). Therefore, a new meta-analysis of IF-AT studies is thus in place to identify the sources of this inconsistencies.

Purposes of the study

The purposes of this study were three-fold. Firstly, it aimed to assess whether the implementation of IF-AT in a multiple-choice test is more effective for students to improve their learning than traditional multiple-choice test without immediate feedback. Secondly, it attempted to assess the relative effectiveness between the computer-based form and the scratch-and-win IF-AT. Thirdly, it tried to explore possible moderating effects (sources of inconsistencies) on IF-AT tests in case of heterogeneity of effect sizes.

Introduction of IF-AT

IF-AT is a modified multiple choice. It provides instant corrective feedback to learners during testing (Epstein et al., 2002). Accordingly, students know immediately if their answer is correct or not. In the case of an incorrect answer, they will have an opportunity to supply a second attempt and receive a partial credit for a correct second try. IF-AT is also called an answer-until-correct (AUC) test introduced by Pressey in 1926, designed to improve effectiveness of learning (Kulik & Kulik, 1988). Based on the format of multiple choices, IF-AT allows students to continue answering if they miss the correct option, and the procedure of answering is not terminated until they make a right choice.

There are several aspects of IF-AT to lay claim to its unique features, including immediate feedback, self-learning while testing, less test anxiety, longer retention and partial knowledge recognition (DiBattista, 2005). Pressy (1926) and Skinner (1954) argued that students' learning was often interrupted by long intervals, thus they suggested that teachers should adopt immediate feedback to reinforce what students have just learned in class. According to Thorndike's law of effect, a positive feedback intervention should be immediately given as reinforcement (Kluger & DeNisi, 1996). In the IF-AT model, students are able to receive immediate feedback for correct answers, which may verify and reinforce what students have learned. Gauging students' partial knowledge is one of the major benefits of IF-AT as students are not totally denied by choosing a wrong answer at the first time. Instead, a wrong answer will prompt them to have a further opportunity to organize their knowledge to find the right answer (Brosvic et al, 2006). Students are also more motivated by the fairness of IF-AT because their partial knowledge is recognized, which makes IF-AT more attractive than the scantron form (DiBattista & Gosse, 2006).

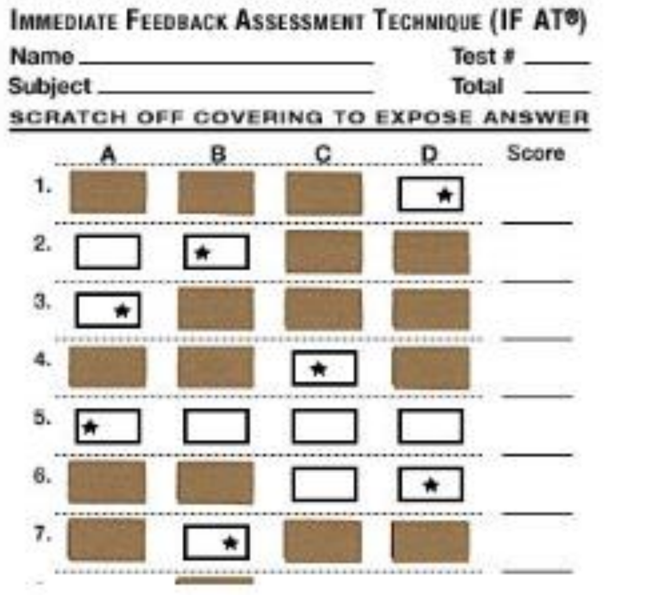
Development of IF-AT

In the early stage, Pressey (1926) designed an answering machine with four numbered buttons for students to answer after reading a multiple-choice question. This machine would not display the next question until students provided the correct answer (Thompson, 1975). Due to its inconvenience, Pressey later used punch boards with pencils for an IF-AT test.

After 1970, there were two variations of the IF-AT test, one with a latent image transfer sheet, the other one with a carbon coating sheet (Thompson, 1975; Anderson, Kulhavy, & Andre, 1971). With a latent image transfer sheet, a special chemical is used to cover the answering options in a multiple-choice testing item and the student uses a special pen to answer the question in order to expose its correctness. A carbon coated IF-AT test is similar to the previous one, except the options were shielded by carbon and requires the student to erase carbon on their selected answer. The use of chemicals and carbon coating on test sheets were still not easy enough to use an IF-AT test in a classroom. Furthermore, an IF-AT test was much more expensive than the traditional multiple-choice test, which was a reason why not many IF-AT studies were conducted.

Difficulties could not be overcome in IF-AT tests until technology had made a huge improvement and the use of personal computers on campus had become more prevalent. Advanced computer technology made intervening feedback more effective and consequently more studies adopted the computer interface.

In 2001, Epstein, Epstein, and Brosvic developed an innovative IF-AT form using a card with an opaque film on the multiple-choice options and allowing the student to scratch off an option as if they were scratching a lottery ticket. If a scratched option indicated a star in a rectangle, it meant that a student had made a correct response. Scratched answers without a star indicated that students have not selected the correct answer (see Figure 1). They could not stop answering until they had found the correct answer for the question. For a multiple-choice test item with four options, a student would receive full credit when a star appeared in a scratched option on the first try, partial credit would be given for their second or third try. Although many studies have discussed the effectiveness of the scratch-card IF-AT, there have not been any meta-analytic synthesis reported comparing its relative effectiveness with traditional multiple-choice test without instant feedback.



IMMEDIATE FEEDBACK ASSESSMENT TECHNIQUE (IF AT®)

Name ____________________ Test # ____

Subject ____________________ Total ____

SCRATCH OFF COVERING TO EXPOSE ANSWER

	A	B	C	D	Score
1.	▨	▨	▨	★	____
2.	□	★	▨	▨	____
3.	★	▨	▨	▨	____
4.	▨	▨	★	▨	____
5.	★	□	□	□	____
6.	▨	▨	□	★	____
7.	▨	★	▨	▨	____

Figure 1 IF-AT scratch card, Epstein Enterprises (2017)

Method

Identifying Primary Studies

To retrieve relevant studies of IF-AT, a systematic and comprehensive search was conducted on several electronic databases: Google scholar, ERIC, Science Direct, ProQuest, and Springer Link. A total of 563 articles were located from the searches. Key words or a combination of key words used in the search included immediate feedback, assessment, technique, answer, and until correct. Further screening by reviewing abstracts of articles, 147 articles were retained. After reading each of the 147 full-text articles, there were 45 English-language studies eligible for the purpose of extracting the effect size for the standardized mean difference between the IF-AT experimental form and the multiple-choice control form. Table 1 reports a detailed summary including effect sizes, sample sizes, study quality, and major study characteristics from each of the 45studies included in the initial review.

Table 1

Summary of Studies Included & Excluded from the Screening Process

Study name	Effect size (Hedge's g)	Subjects	Exp group N	Ctrl group N	IF-AT form	Testing domain	Sample size	Exp design	Impact factor	Reliability
1970 Heald*	0.792	0	18	18	2	0	1	0	Doctor Dissertation	N/A
1971 Olson*	0.367	0	51	50	2	0	3	0	Doctor Dissertation	N/A
1975 Carel*	0.665	0	18	18	2	1	1	0	Doctor Dissertation	0.9
1975 Thompson*	0.329	0	15	21	2	1	1	0	Doctor Dissertation	0.86
1976 Kulhavy et al.*	0.563	0	30	30	2	1	3	0	4.24	N/A
1976 O'neil*	1.606	0	28	31	2	2	2	1	1.208	0.88
1990 Clariana	0.924	0	50	99	0	0	3	1	N/A	0.56
1991 Clariana et al.	0.618	1	10	10	0	0	1	0	0.725	0.85
1991 Pridemore & Klein	1.084	0	45	45	0	0	3	0	0.725	0.69
1995 Morrison et al.	0.457	0	20	19	0	0	2	0	2.877	0.76
2000 Clariana et al.	0.529	1	27	27	0	0	2	0	0.725	0.76
2001 Corbett & Andserson	0.940	0	30	30	0	0	3	1	Conference	N/A
2003 Samuel & Wu 2003	1.006	2	28	39	1	0	3	1	N/A	0.88
2003a Epstein et al. ,Study 2*	2.226	3	60	60	1	0	3	0	0.86	N/A
2004 Dihoff Brosvic Epstein	1.083	0	32	48	1	0	3	0	0.86	N/A
2004 Jones	0.656	0	32	32	0	0	3	0	Doctor Dissertation	N/A
2004 Rosa Leow	0.901	0	18	17	0	0	1	0	1.745	N/A
2006 Brosvic Dihoff Epstein	1.283	4	40	40	1	1	3	1	0.86	0.96
2006 DiBattista & Gosse	0.428	0	154	154	1	1	3	1	1.653	0.85
2007 Butler et al.	0.902	0	46	47	0	0	3	0	3.224	N/A
2007 Murphy	0.491	0	108	107	0	0	3	0	2.29	N/A
2008 Fitch & Hulgin	0.596	2	30	95	0	1	3	1	0.844	0.88
2008 Persky et al.	0.246	0	144	132	1	1	3	1	1.109	N/A
2009 Carmichael	0.479	0	200	200	1	1	3	1	0.81	N/A
2009 Dibattista et al.	0.966	0	132	132	1	0	3	0	1.653	0.7
2010 Finn & Metcalfe	0.652	0	24	24	0	0	2	0	2.253	N/A
2010 Prieto et al.	0.546	0	47	47	0	1	3	1	0.31	0.88
2011 Bowman et al.	0.128	0	12	11	1	1	1	0	2.341	N/A
2011 Parisa Razagifard	0.938	0	30	30	0	0	3	0	0.435	N/A

2011 Simley CIFAT	0.550	0	50	49	0	1	3	1	Master thesis	N/A
2011 Simley IFAT	0.524	0	50	49	1	1	3	1	Master thesis	N/A
2012 Marsh et al.	0.699	0	28	28	0	0	2	1	2.118	N/A
2012 Valdez	0.225	0	22	22	0	0	2	0	0.725	0.68
2013 Butler Marsh	0.858	0	20	20	0	0	2	0	5.24	0.89
2013 Kehrer et al.	0.368	1	61	61	0	1	3	0	Conference	N/A
2013 Peck et al*	1.349	0	56	62	1	2	3	1	0.91	N/A
2013 Slepkov Shiell	0.493	0	131	129	1	1	1	1	0.956	0.82
2014 Ghani	0.424	0	20	20	0	1	1	0	Conference	0.6
2014 Schneider et al.	0.320	0	349	151	1	1	3	1	Conference	0.8
2015 Farland et al.	0.150	0	30	54	1	1	3	1	1.68	0.94
2015 Maurer & Kropp	1.245	0	85	91	1	0	3	1	0.19	N/A
2015 Merrel et al.	0.129	0	44	44	1	1	3	1	N/A	N/A
2015 Mohrweis & Shinham	0.687	0	100	73	1	0	3	0	N/A	N/A
2015 Peters*	0.076	5	43	33	1	0	3	1	Doctor Dissertation	0.96
2016 Slepkov	0.597	0	414	353	1	1	3	1	1.419	0.83

Note. IF-AT form: 0=Computer-based 1=Scratch-card 2=Latent image or carbon coating
Testing domain: 0=Social sciences 1=Math & science 2=Mixed testing material
Sample size/per group: 1= <20 2= 20~29 3= ≧ 30
Subjects: 0=Undergraduates 1=High school 2=Primary school 3=Pre-school 4=Students with difficulty in learning math 5=Newly Arabian immigrants
Experiment design: 0=Randomly assigned 1=Quasi-experiment
*A study name with a star indicates an excluded study during the screening process.

A total of 59 effect sizes were initially extracted from 45 studies, some of them were dependent due to repeated measures. The synthesis of dependent effect sizes in a study was required. Through sensitivity analysis, there were 5 studies removed from the analyses due to their unusual effect sizes. For instance, the Epstein et al. (2003) study contributed huge heterogeneity due to the use of pre-school young kids (with an unusually large effect size). Similarly, the Peters (2015) study and the Brosvic et al. (2006) study also contributed huge heterogeneity because their study subjects were Arabian adult immigrants (with an unusually small effect size) and students with learning difficulties in mathematics (with an unusually large effect size). Subjects in the above three studies were idiosyncratic from the other studies. In addition, testing domains used in two studies (Peck, Werner, & Raleigh, 2013; O'Neil, Razor & Bartz, 1976) were mixed so both studies were removed as well. After the filtering process, there were 40 studies left, 5 IF-AT (AUC) studies used latent image sheets and carbon coatings sheets, 20 IF-AT studies used computer-based testing, and 15 IF-AT studies used a scratch-card form. Five studies of IF-AT with latent image and carbon coating were obsolete and inconvenient for use. After the removal of these five studies, there were 35 studies left for this current meta-analysis (see Figure 2).

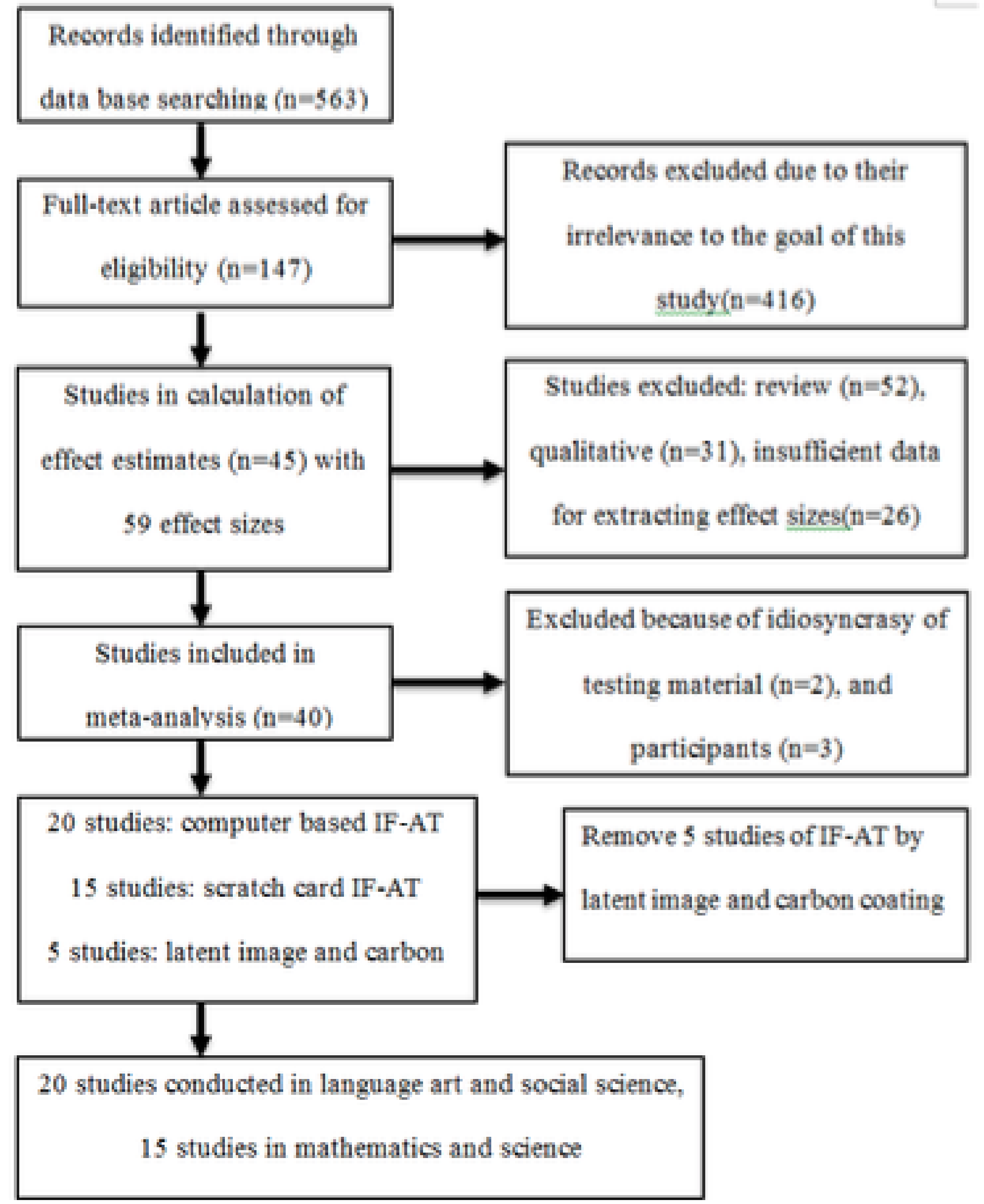

Records identified through data base searching (n=563)
Full-text article assessed for eligibility (n=147)
Records excluded due to their irrelevance to the goal of this study(n=416)
Studies in calculation of effect estimates (n=45) with 59 effect sizes
Studies excluded: review (n=52), qualitative (n=31), insufficient data for extracting effect sizes(n=26)
Studies included in meta-analysis (n=40)
Excluded because of idiosyncrasy of testing material (n=2), and participants (n=3)
20 studies: computer based IF-AT
15 studies: scratch card IF-AT
5 studies: latent image and carbon
Remove 5 studies of IF-AT by latent image and carbon coating
20 studies conducted in language art and social science,
15 studies in mathematics and science

Figure 2 Flow chart of literature search for IF-AT meta-analysis

Among the 35 IF-AT studies, some of them utilized reading tests to assess students' reading achievement in vocabulary and reading comprehension. The content of those reading tests was different from school quizzes or examinations (Clariana, 1990). Results of these studies indicated that IF-AT could facilitate a student's memory retention. On the other hand, the other IF-AT studies were used to replace the multiple-choice test to administer quizzes or examinations at schools. They were applied, for example, in physics (Slepkov, 2013), chemistry (Merrel et al., 2015), biology (Carmichael, 2009), pharmacy (Persky, 2008), psychology (Smiley, 20011), and commerce (Mohnweis a& Shinham, 2015).

Coding Procedures

The first author, guided by the second author with more required necessary training and experiences in coding studies for a meta-analysis, coded all of the studies in the current meta-analysis. The following information was coded from each article identified for inclusion in the meta-analysis: effect sizes, sample sizes, sample characteristics, reliability indices of measures, and moderator variables (such as testing forms, participants, and testing domains) as shown in Table 1. All articles were coded by the first author of this paper and rechecked by the second author. When disagreements occurred, these were resolved through mutual discussion between us and reached a consensus.

Effect Size Calculation

The standardized effect size (Cohen's *d*) is the mean difference (between the experimental group and the control group) divided by the pooled standard deviation of the two groups. If only the F or t value reported in a study, a calculator developed by Lipsy and Wilson (2001) was used to convert it into a Cohen's *d*. Effect sizes of some studies reporting η^2, ω^2 , and Mann-Whitney *U* were converted into a Cohen's *d* using a formula provided by Rosnow, Rosenthal and Rubin (2000). As for researchers that only reports a two-way ANOVA, Lipsy and Wilson's (2001) effect size converter was used to compute effect sizes. But, it is necessary to include relevant information such as mean square error, degree of freedom, F values of interaction and main factors. The calculation of effect-size variance in a study is given by Formula (1):

$$\frac{Ne+Nc}{Ne\times Nc}+\frac{d^2}{2(Ne+Nc)}$$

Where Ne is the number of subjects in an IF-AT group, Nc is the number of subjects in a traditional multiple-choice group. The thirty-five studies collected in this research were conducted in experimental research study with an experimental group (IF-AT) and a control group (traditional multiple-choice testing). However, there were 48 effect sizes extracted from 35 studies in this IF-AT meta-analysis. Thus, some studies involve more than one effect size on IF-AT. There is a concern about the dependence of each effect size in a repeated measures study, which may cause inflated weighting when synthesizing effect sizes into a pooled mean (Li, 2015). The modified variance is calculated by Formula (2) (Borenstein et al. 2009, Li, 2015).

$$\mathrm{Var}(1/\mathrm{m}\sum_{i=1}^{m} Yi\ )=(1/\mathrm{m})^2(\sum_{i=1}^{m} Vi + 2\sum_{i\neq j}\left(rij\sqrt{Vi} * \sqrt{Vj}\right)$$

Besides, a calculated Cohen's *d* may be inflated due to small samples in some studies (Hedges and Olkin, 1985), a conversion of Cohen's *d* to Hedge's *g* is required. The standard mean difference in Hedge's *g* can be calculated using Formula (3).

$$\mathrm{g} = \left(1-\frac{3}{4N-9}\right)d$$

Software Used and Number of Studies Needed

All data analyses in this study were conducted by Comprehensive Meta-Analysis software (Borenstein, Hedges, Higgins, & Rothstein, 2009) for estimation of pooled effect sizes, Q statistics, and confidence intervals. An effect size synthesizer, an EXCEL add-in Macro (Li, 2015), was also used to estimate the optimal number of studies needed for a meta-analysis with a sufficient statistical power at .80. Assuming the significant level of alpha is set at .05, the average number of participants in the experimental and the control groups is 20, and the minimum effect size of mean difference is given at .20, which is the smallest effect cut-off criterion suggested by Cohen (1992). As the result indicated, at least 14 studies are required for a fixed model and 22 studies are required for a random model to meet the sufficient statistical power at .80 with a medium degree of heterogeneity.

Fixed-Effects Models versus Random-Effects Models

To report the synthesized effect size, either the fixed model or the random model can be adopted. Each takes up a different underlying assumption. The fixed model assumes that there exists a true effect size among studies. The difference of each observed effect size is just due to sampling error (Borenstein, 2009). Under the random model, it assumes that the true effect size can vary from study to study. The different effect size of each observed study is caused both by participants and interventions (Borenstein, 2009). Since IF-AT studies have been performed on diverse subjects and populations, it would be more justified in the current study to interpret the summary data based on a random-effects model.

Between-study Heterogeneity Analysis

Three indices that can be used to judge heterogeneity in the study results are Q statistic, τ^2 index and I^2 index. Nevertheless, Q only suggests the presence of heterogeneity. It is better to include τ^2 and I^2 to quantify the degree of heterogeneity in a meta-analysis. When τ^2 is less than .04, it could reveal a low degree of heterogeneity in a meta-analysis. When τ^2 is between .04 and .14, studies in a meta-analysis are regarded as s moderate degree of heterogeneity; and there could be a substantial heterogeneity in a meta-analysis when τ^2 is over (Spiegelhalter, Abrams, & Myles, 2004). When I^2 is between 0 and 25%, it is considered as trivial or there is no heterogeneity among studies. When I^2 is above 75%, it is considered to have a high degree of heterogeneity among studies. When I^2 is between 25% and 75%, it is considered to have a medium degree of heterogeneity among studies (Higgins,Thompson, Deeks, & Altman, 2003). To identify the sources of heterogeneity is one of the most important goals of meta-analysis. In case of severe heterogeneity, a moderator analysis will be conducted to explore the possible sources of heterogeneity.

Publication Bias Analysis with a Covariate

Without a covariate, the current IF-AT meta-analysis was robust to the file drawer problem (Rosenthal, 1979). From the funnel asymmetry test**,** it appeared that the funnel plot was symmetric for the random model (see Figure 3). In addition, Eggers' FAT regression test (Sterne, Becker, & Egger, 2005) also revealed an insignificant publication bias effect with PB=.968, t value=1.42 (p=.166). Since asymmetry was often not caused by publication bias alone, the application of trim and fill would not be appropriate. The asymmetry might be due to other causes such as heterogeneity of effect sizes. Thus, the FAT-PET meta-regression with a covariate seems to be a better way to correct effect size estimates for publication bias and to investigate various sources of heterogeneity. This approach first conducts a funnel-asymmetry testing (FAT), a test for publication bias and then conducts the precision-effect test (PET) to examine whether there is a true effect beyond publication bias (Stanley, 2008; Doucouliagos, Stanley & Viscusi, 2014). If there is a non-zero effect and no substantial heterogeneity, PEESE (Precision-Effect with Standard Error) can be used to obtain the net effect size by correcting heterogeneous bias beyond publication bias and moderating effect (Stanley, 2017; see also the FAT-PET- PEESE section latter).

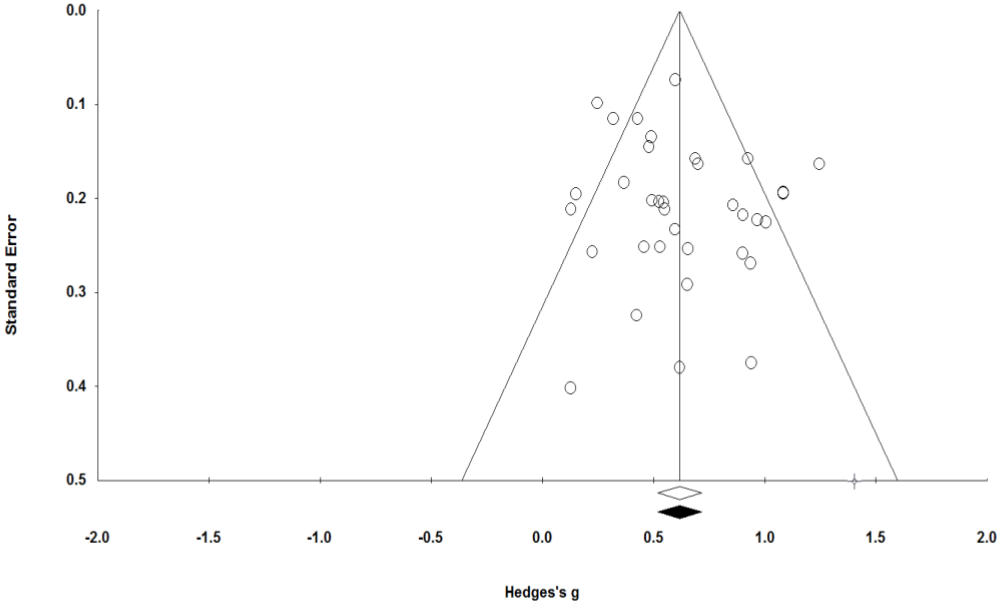
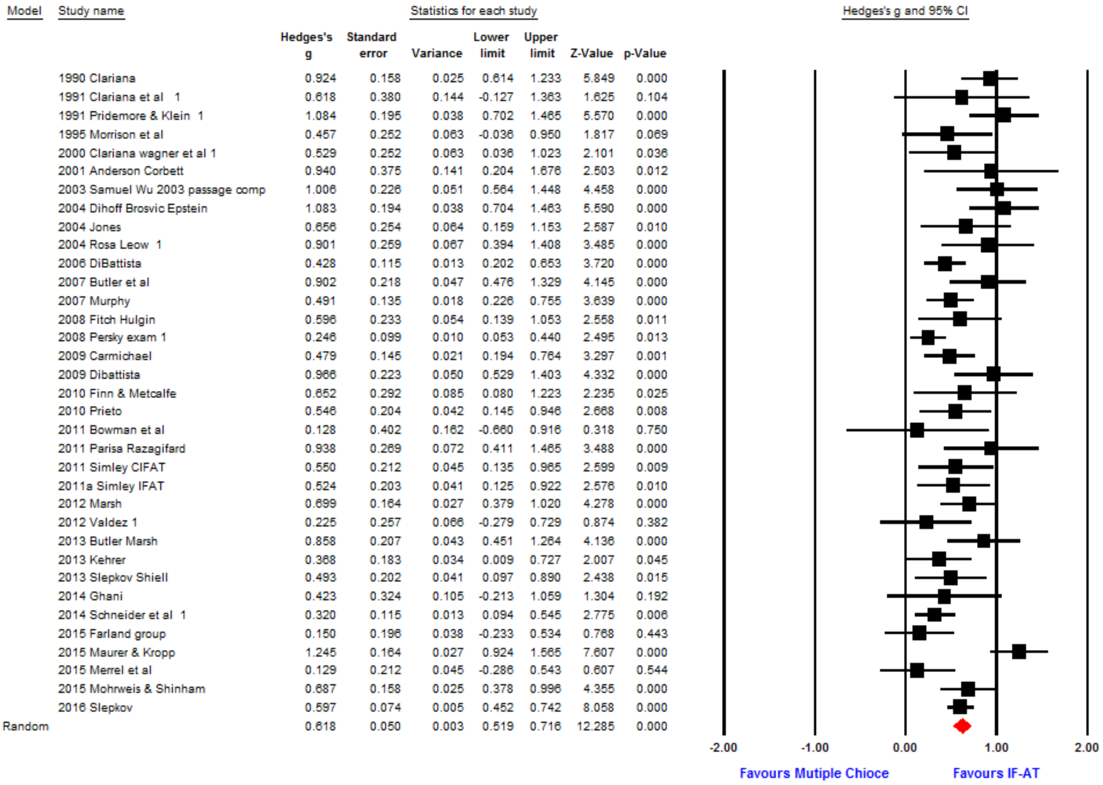


Figure 3 A symmetric funnel plot of the effect sizes for the 35 IF-AT studies(random)

The Overall Effectiveness of IF-AT and Heterogeneity

Currently IF-AT studies are either computer-based or scratch card. It is reasonable to proceed with our meta-analysis based on these 35 IF-AT studies. The statistical power of 35 studies was .98 for the fixed model and .89 for the random model. It showed that the number of studies was sufficient to detect the hypothesized effect size (Li, 2015). In the 35 recent IF-AT studies, there were 2623 participants in the experimental group of IF-AT and 2489 participants in the multiple choices control group. According to these 35 studies of the computer-based IF-AT and the scratch-card IF-AT, the overall effect size was .581, CI:[.524, .639] in the fixed model, and .618, CI: [.519, .716] in the random model(See Table 2). The effect size was not far away from the previous meta-analysis result (.53) by Bangert-Drowns et al. (1991). It was also larger than the benchmark of the feedback-intervened mean difference (.41) by Kluger and DeNisi (1996).

The forest plot as shown in Figure 4 offers a simple visual representation of the information of individual and overall effect sizes and variation of effect sizes between studies. All 35 effect sizes, along with the 95% confidence, are portrayed graphically in Figure 4. Among the 35 studies, effect sizes of 28 studies were statistically significant because their confidence intervals did not include zero (no effect), indicating that their effects significantly differ from zero. To assess heterogeneity of effect-size estimates from individual studies, the obtained Q value was 84.567(5.081+79.486) with p<.001, I^2=59.80%, and τ^2=.047, which revealed moderate heterogeneity. It might need a moderator or moderators to identify sources of heterogeneity.



Model	Study name	Hedges's g	Standard error	Variance	Lower limit	Upper limit	Z-Value	p-Value
	1990 Clariana	0.924	0.158	0.025	0.614	1.233	5.849	0.000
	1991 Clariana et al 1	0.618	0.380	0.144	-0.127	1.363	1.625	0.104
	1991 Pridemore & Klein 1	1.084	0.195	0.038	0.702	1.465	5.570	0.000
	1995 Morrison et al	0.457	0.252	0.063	-0.036	0.950	1.817	0.069
	2000 Clariana wagner et al 1	0.529	0.252	0.063	0.036	1.023	2.101	0.036
	2001 Anderson Corbett	0.940	0.375	0.141	0.204	1.676	2.503	0.012
	2003 Samuel Wu 2003 passage comp	1.006	0.226	0.051	0.564	1.448	4.458	0.000
	2004 Dihoff Brosvic Epstein	1.083	0.194	0.038	0.704	1.463	5.590	0.000
	2004 Jones	0.656	0.254	0.064	0.159	1.153	2.587	0.010
	2004 Rosa Leow 1	0.901	0.259	0.067	0.394	1.408	3.485	0.000
	2006 DiBattista	0.428	0.115	0.013	0.202	0.653	3.720	0.000
	2007 Butler et al	0.902	0.218	0.047	0.476	1.329	4.145	0.000
	2007 Murphy	0.491	0.135	0.018	0.226	0.755	3.639	0.000
	2008 Fitch Hulgin	0.596	0.233	0.054	0.139	1.053	2.558	0.011
	2008 Persky exam 1	0.246	0.099	0.010	0.053	0.440	2.495	0.013
	2009 Carmichael	0.479	0.145	0.021	0.194	0.764	3.297	0.001
	2009 Dibattista	0.966	0.223	0.050	0.529	1.403	4.332	0.000
	2010 Finn & Metcalfe	0.652	0.292	0.085	0.080	1.223	2.235	0.025
	2010 Prieto	0.546	0.204	0.042	0.145	0.946	2.668	0.008
	2011 Bowman et al	0.128	0.402	0.162	-0.660	0.916	0.318	0.750
	2011 Parisa Razagifard	0.938	0.269	0.072	0.411	1.465	3.488	0.000
	2011 Simley CIFAT	0.550	0.212	0.045	0.135	0.965	2.599	0.009
	2011a Simley IFAT	0.524	0.203	0.041	0.125	0.922	2.576	0.010
	2012 Marsh	0.699	0.164	0.027	0.379	1.020	4.278	0.000
	2012 Valdez 1	0.225	0.257	0.066	-0.279	0.729	0.874	0.382
	2013 Butler Marsh	0.858	0.207	0.043	0.451	1.264	4.136	0.000
	2013 Kehrer	0.368	0.183	0.034	0.009	0.727	2.007	0.045
	2013 Slepkov Shiell	0.493	0.202	0.041	0.097	0.890	2.438	0.015
	2014 Ghani	0.423	0.324	0.105	-0.213	1.059	1.304	0.192
	2014 Schneider et al 1	0.320	0.115	0.013	0.094	0.545	2.775	0.006
	2015 Farland group	0.150	0.196	0.038	-0.233	0.534	0.768	0.443
	2015 Maurer & Kropp	1.245	0.164	0.027	0.924	1.565	7.607	0.000
	2015 Merrel et al	0.129	0.212	0.045	-0.286	0.543	0.607	0.544
	2015 Mohrweis & Shinham	0.687	0.158	0.025	0.378	0.996	4.355	0.000
	2016 Slepkov	0.597	0.074	0.005	0.452	0.742	8.058	0.000
Random		0.618	0.050	0.003	0.519	0.716	12.285	0.000

Figure 4 Forest plot of the effect sizes for the 35 IF-AT studies

Computer-based IF-AT and Scratch-card IF-AT

There remained a controversial issue regarding the effectiveness of these two forms of IF-AT. Some researchers have concluded that the scratch-card IF-AT was superior to the computer-based IF-AT (Epstein et. al, 2002), while others argued that there is no mean difference of effect sizes between them (Smiley, 2011). In this research, a sub-group analysis was applied to compare the mean difference between both IF-AT forms. The results of sub-group analysis in Table 2 shows that the overall effect size for the computer-based IF-AT is .668 in the fixed model and .669 in the random model. For the scratch-card IF-AT, it is .531 in the fixed model and.568 in the random model. The $Q_{between}$ statistic is 5.081 (p value=.024), indicating a significant mean difference between both forms of IF-AT. It seems evident that computer-based IF-AT is more effective; however, this result is not appropriate because the Q_{within} statistic is 79.486 (p<.001), revealing that the remaining heterogeneity was still too much to be ignored. Consequently, it is not adequate to combine the studies for an average effect, because the mean effect size difference between the computer-based IF-AT and the scratch-card IF-AT may moderate with other factors such as testing domain or testing form.

The Q values in Table 2 reveals that the 20 computer-based IF-AT studies are homogeneous (Q=21.275, p=.332; I^2=10.695%), while the 15 studies of the scratch-card IF-AT studies are heterogeneous (Q= 58.211, P<.001; I^2 = 75.95 %). As a result, it entails a sub-group analysis to find its potential moderator(s) for explaining the heterogeneity left within 15 studies of the scratch-card IF-AT (see Table 5 for details).

Table 2

Sub-group Comparison between Scratch Card and Computer Based IF-AT

Assessment form	**Mean effect size**	**Confidence interval**	**Heterogeneity**	**P-value for Q**

Scratch card
15 studies
	.531(F)
	.568(R)
		.458-.603(F)
		.408-.728(R)
			Q=58.211;I^2=75.950%
			τ^2=.069
				P<.001

Computer based 20 studies	.668(F)
	.669(R)
		.573-.764(F)
		.567-.772(M)
			Q=21.275; I^2=10.695%
			τ^2=.006
				P=.332

Total	.581(F)
	.618(R)
		.524-.639(F)
		.519-.716(M)
			$Q_{between=}$5.081
			$Q_{within=}$79.486
			Q_{total}=84.567; I^2=59.8%
			τ^2=.047
				P=.024
				P<.001
				P<.001

Moderator Analysis by Test Material

A good moderator is able to explain the main source of heterogeneity between studies, beyond the variability due to sampling error. A bare-bones analysis of the 35 studies indicated that the percentage of variance accounted for by the experimental artifact was 40.65 %. Hunter and Schmidt (2004) suggested that if it is lower than 75%, a moderator analysis is required to explain the heterogeneity except for the sampling error. Therefore, three potential moderators selected for this IF-AT meta-analysis are types of participants, testing domains and testing forms. A valid moderator ought to satisfy that the $Q_{between}$ statistic should be significant and the Q_{within} statistic should be nonsignificant. The initial analysis of meta-regression (See Table 3) found that testing domain was the most significant moderator ($Q_{between}$ =39.311 p <.001), which explained 46.48% of heterogeneity of IF-AT.

Table 3

Meta Regression for Moderator Analysis Based on 35 IF-AT Studies

Moderators	$Q_{between}$	Q_{within}	R^2
Testing forms	5.081 (p=.024)	79.486 (p<.001)	6.01%
Type of Participants	.4362 (p=.509)	84.131 (p<.001)	.52%
Testing domains	39.311 (p<.001)	45.256 (p=.076)	46.48%

Accordingly, a sub-group analysis by testing domain shown in Table 4 revealed that the overall effect size of 20 IF-AT studies conducted in language arts and social sciences (LS) was .806 in the fixed model (.807 in the random model), and that of 15 IF-AT studies conducted in mathematics and science was (MS) .430 in the fixed model (.421 in the random model). Obviously, IF-AT studies conducted in language arts and social sciences were more effective than those conducted in mathematics and science ($Q_{between}$=39.311, p<.001). Since testing domain has explained 46.48% of total heterogeneity, heterogeneity within studies dropped sharply (Q_{within} =45.256, p=.076). In the LS group, the Q_{within} value was 29.465 (p=.059), I^2=35.516%, and τ^2=.024; while the Q_{within} value was 15.792 (p=.326), I^2=11.348%, and τ^2=.003 in the MS group (See Table 4). Therefore, there is no need to continue a sub-group analysis.

Table 4

35 IF-AT Studies Moderated by Testing Domains

Testing domain	**Mean effect size**	**Confidence interval**	**Heterogeneity**	**P-value for Q**

Language arts & Social sciences
20 studies (LS)
	.806(F)
	.807(R)
		.715-.897(F)
		.690-.924(R)
			Q=29.465;I^2=35.516%τ^2=.024	P=.059

Math & Science
15 studies (MS)
	.430(F)
	.421(R)
		.355-.505(F)
		.338-.505(R)
			Q=15.792;I^2=11.348%τ^2=.003	P=.326

Total			$Q_{between}$:39.311 Q_{within}:45.257	P<.001 P=.076

More specifically, 35 studies can be further broken down by IF-AT form, as shown in Table 5 and Table 6. As a successful moderator, the variance of heterogeneity between the 15 scratch-card IF-AT studies (Q=58.211, p<.001, see Table 5) was explained largely by testing domain. There were five scratch-card IF-AT studies conducted in language arts and social sciences and 10 scratch-card IF-AT studies conducted in mathematics and science. Table 5 displays the results of sub-group analysis by testing domain for the scratch-card IF-AT studies. As for total heterogeneity of the 15 scratch-card IF-AT studies, the $Q_{between}$ statistic was 37.400 (p<.001) and the Q_{within} statistic was 20.811 (p=.077), which indicates testing domain did explain 64.25% (37.400/58.211) of total heterogeneity in the scratch-card IF-AT studies. The overall effect of the scratch-card IF-AT in language arts and social sciences was .988 in the fixed model (.992 in the random model), and that of scratch-card IF-AT in mathematics and science was .419 in the fixed model (.392 in the random model). Indicated by the significant $Q_{between}$ statistic (Q=37.40, p<.001), it can be inferred that scratch-card IF-AT tests have worked significantly better in language arts and social sciences than in mathematics and science.

Table 5

Scratch Card-based IF-AT Moderated by Testing Domain

Testing domain	Mean effect size	Confidence interval	Heterogeneity	P-value for Q

**Language arts & Social sciences
5 studies (LS)**
	.988(F)
	.992(R)
		.824-1.151(F)
		.783-1.201(R)
			Q=6.361; I^2=37.114%
			τ^2=.021
				P=.174

**Mathematics & science
10 studies (MS)**
	.419(F)
	.392(R)
		.338-.500(F)
		.279-.504(R)
			Q=14.451;I^2=37.719%;
			τ^2=.011
				P=.107

Total			$Q_{between=}$37.400 $Q_{within=}$20.811 Q_{total}=58.211	P<.001 P=.077 P<.001

As to the relative effectiveness between the computer-based IF-AT and the scratch-card IF-AT as shown in Table 6, their effect sizes interacted significantly with testing domains. As for total heterogeneity of the 20 computer-based card IF-AT studies, the $Q_{between}$ statistic was 4.175 (p=.041) and the Q_{within} statistic was 17.101 (p=.516), which indicated testing domains explain 19.62% (4.175/21.275) of total heterogeneity in the computer-based IF-AT studies. The overall effect size of the 15 computer-based IF-AT studies conducted in language arts and social sciences was .725 in the fixed model and .728 in the random model, and the overall effect size of 5 computer-based IF-AT studies conducted in mathematics and science was .494 in the fixed and random models. As indicated by the significant $Q_{between}$ statistic (Q=4.175, p=.041), the computer-based IF-AT worked more effectively in language arts and social sciences than in mathematics and science.

Table 6

Computer- Based IF-AT Moderated by Testing Domain

Testing domain	Mean effect size	Confidence interval	Heterogeneity	P-value for Q

Language arts & Social sciences 15 studies (LS)	.725(F)
	.728(R)
		.615-.834(F)
		.607-.848(R)
			Q=16.255;I^2=13.873%
			τ^2=.008
				P=.298

**Mathematics & science
5 studies (MS)**
	.494(F)
	.494(R)
		.301-.686(F)
		.301-.686(R)
			Q=.845; I^2=0.00%;
			τ^2=.00
				P=.932

Total			$Q_{between=}$4.175
			$Q_{within=}$17.100
			Q_{total}=21.275
				P=.041
				P=.516
				P=.332

Based on the $Q_{between}$ statistics shown in Tables 5 and 6, it seems reasonable to conclude that IF-AT tests, regardless of scratch-card or computer-based, perform better in the context of language arts and social sciences than in the context of mathematics and science. Additionally, most computer-based IF-AT studies were conducted in language arts and social sciences (15/20), while most scratch card IF-AT studies were conducted in mathematics and science (10/15). This might explain why the computer-based IF-AT worked more effectively than the scratch-card IF-AT (.668 versus .531 in the fixed model, $Q_{between}$=5.081 and p=.024; see Table 2).

Table 7 summarizes the data from Tables 5 and 6. In the context of LS, the scratch-card IF-AT (overall effect size=.988 in the fixed model and .992 in the random model) outperformed the computer-based IF-AT (the overall effect size=.725 in the fixed model and .728 in the random model) as indicated by the significant $Q_{between}$ statistic (6.849, p=.009). In the context of MS, it seemed that the computer-based IF-AT (the overall effect size=.494 in the fixed and random models) appeared as equally effective as the scratch-card IF-AT (overall effect size=.419 in the fixed model and .392 in the random model) as indicated by the insignificant $Q_{between}$ statistic (.497, p=.481). Though research findings of Epstein et al. (2002) and Smiley (2011) have contradicted each other, our results may render both of them feasible. As a matter of fact, testing domain moderated the relative effectiveness between the scratch-card IF-AT and the computed-based IF-AT . When IF-AT tests were conducted in language arts and social sciences, the scratch-card IF-AT was more effective than the computer-based IF-AT, and this result is consistent with Epstein et al. (2002) study. When IF-AT was conducted in the MS context, there were no significant difference between studies of the scratch-card IF-AT and the computer-based IF-AT. This result is consistent with the finding reported by Smiley (2011). Based on data for the fixed effect model on Table 7, an interaction plot can be created in Figure 5.

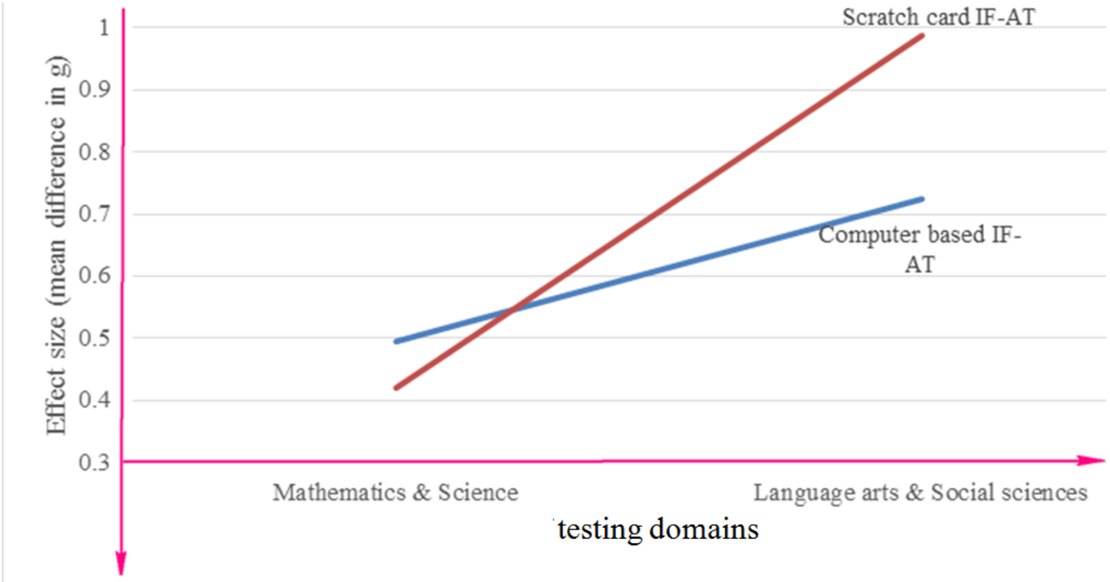


Figure 5 Mean effects interact with forms of IF-AT and types of testing domains

It also demonstrates that the effectiveness between the scratch-card IF-AT and the computer-based IF-AT depended on testing domain. The mean effect of IF-AT interacts with its forms and testing domain because the effectiveness between the scratch-card IF-AT and the computer-based IF-AT in the context of language arts and social sciences differed from the one in the context of mathematics and science.

Table 7

Comparison between Scratch Card-based IF-AT and Computer-based IF-AT by Testing Domain

Testing domain	**Scratch card-based vs Computer-based**		**$Q_{between}$**
	(15 studies) vs	**(20 studies)**	
Language arts & Social sciences	.988(F)	.725(F)	6.849
20 studies (LS)	.992(R)	.728(R)	(p=.009)
# of studies	5	15	

Mathematics & Science
15 studies (MS)
# of studies 	
	.419(F)
	.392(R)
	10
		.494(F)
		.494(R)
		5
			.497
			(p=.481)

Analysis of FAT-PET- PEESE Meta-Regression

In a traditional meta-analysis, the effect sizes are assumed to be homogeneous. If not, sub-group or meta-regression analysis is required to identify sources of heterogeneity except for sampling error. The FAT-PET-PEESE meta-regression provides a promising tool for accommodating publication bias and moderating effect to calibrate the overall net effect sizes (in case of them being heterogeneous) when there is evidence of non-zero effect. However, the method is not a panacea (Stanley & Doucouliago, 2012).This method estimates more accurately the overall effect sizes only in the fixed model because the FAT-PET-PEESE meta-regression could cause a larger bias in the random model than in the fixed effect model. Stanley (2017) also identified that a poor result may happen in conducting the FAT-PET-PEESE meta-regression under the following scenarios: (1) insufficient studies, (2) small samples for all studies, and (3) high heterogeneity from study to study. Our meta-analysis is free of these three limitations. Consequently, the results of the FAT-PET-PEESE meta-regression could be better estimates than the results from conventional meta-analysis in case of dealing with heterogeneous overall effect sizes. To calculate the net effect of IF-AT studies, the meta-regression takes the standard error as well as a moderator as the independent variable. Table 8 presents the net effect sizes in contrast with the original overall effect sizes calculated previously (shown in parentheses).

Table 8

Net Effect Sizes of IF-AT by FAT-PET-PEESE Meta-regression

Studies	**Context**	**Net effects** **(Original)**	**Publication bias** **PB**	**Moderator effect** **ME**
35 IF-AT studies	Overall effect size (35 studies)	.522 (.581)	PB=.968, t=1.42 (p=.166)	

	Computer-based IF-AT (20 studies)	.627 *
		(.668)
			PB=.878
			Z(p)=.667(.500)
				ME=-.1143
				Z(p)=-1.622(.105)

	Scratch card IF-AT(15 studies)	.513 (.531)		

	IF-AT conducted in LS (20 studies)	.856 *
		(.806)
			PB=1.150
			Z(p)=.922(.356)
				ME=-.4010
				Z(p)=-6.105(p<.001)

	IF-AT conducted in MS (15 studies)	.455 (.430)		

* In the meta-regression model, its intercept is the net effect mean of IF-AT.
LS indicates language arts and social science.
MS indicates mathematics and science.
Significant level α =.05: (Z(p)=1.96)

As can be seen in Table 8, the net overall effect of IF-AT was .522 in the fixed model, which was derived by adjusting the original overall effect size (.581) beyond publication bias. More specifically, the net effect of scratch-card IF-AT was .513 and computer-based IF-AT was .627 in the fixed model. Both net effects were slightly modified by partialling out publication bias and moderator effect. It also indicates that there was no significant difference between both testing forms of IF-AT (Z=-1.622, p=.105), though the net effect size of the computer-based IF-AT was higher than that of the scratch-card IF-AT (.627>.513). The number of studies in the meta-analysis is sufficient, so insignificance won't be caused by insufficient statistical power.

By adjusting publication bias and moderator effect, the net effect size of IF-AT studies conducted in language arts and social sciences is .856 and net effect size of IF-AT studies conducted in mathematics and science is .455 in the fixed model. Testing domain was identified to be a valid moderator (the moderator effect of testing domain =-.4010, Z=-6.105, and p<.001). It indicates that the pooled effect size of IF-AT studies conducted in language arts and social sciences was significantly more effective than those conducted in mathematics and science.

To explore the interaction effect between IF-AT form and testing domain, the FAT-PET-PEESE meta-regression was used to compare the net effects between the computer-based IF-AT and the scratch-card IF-AT under both LS and MS contexts. In the context of language arts and social sciences, the net overall effect of the scratch-card IF-AT is .981 and the one of computer-based IF-AT is .717 in the fixed model as shown in Table 9. Both net effects were only slightly adjusted because of a small publication bias effect. As to a moderator effect on IF-AT forms, there is a significant difference of net effects between these two IF-AT forms in the context of language arts and social sciences (moderator effect=.265, Z=2.571, p=.010). The scratch-card IF-AT is more effective than the computer-based IF-AT in language arts and social sciences. This result is consistent with the previous analysis (See Table 7). In the context of mathematics and science, the net effect size of the computer-based IF-AT studies was adjusted from .494 to .618 due to the correction of publication bias effect (PB effect=-2.573). The net effect size of the scratch-card IF-AT studies in mathematics and science is .463. Though the adjusted net effect of the computer-based IF-AT studies is higher than that of the scratch-card IF-AT studies (.618 vs .463) in the context of MS, its moderator effect indicates that there was no significant difference between the scratch-card IF-AT and the computer-based IF-AT (the moderator effect of IF-AT forms=-.156, Z=-1.267, p=.205). This is also consistent with the previous result (see Table 7 for more details). To conclude, there is a significant interaction between testing forms and testing domains in the current IF-AT meta-analysis (also see its graphic display in Figure 5).

Table 9

Net Effect Sizes of Scratch Card IF-AT and Computer-based IF-AT Moderated by Testing Domain (Based on FAT-PET-PEESE Meta-regression)

Testing domain	Scratch Card IF-AT (15 studies) (Original)***	Computer-based IF-AT (20 studies) (Original)***	Publication effect	Moderator effect

IF-AT conducted in LS(20 studies)	.981*
	(.988)
		.717*
		(.725)
			PB=.174
			Z=.092, p=.927
				ME=.265
				Z=2.571, p=.010

IF-AT conducted in MS(15 studies)	.463**
	(.419)
		.618**
		(.494)
			PB=-2.573
			Z=-1.32, p=.187
				ME=-.156
				Z=-1.267, p=.205

* In the meta-regression model, its intercept is the mean of the net effect of the computer-based IF-AT in the context of LS.
** In the meta-regression model, its intercept is the mean of the net effect of the computer-based IF-AT in the context of MS.
*** Values in a parenthesis indicate a corresponding value of an effect size taken from Table 7.

Summary and Discussion

This study adopts a meta-analysis approach to comprehensively and systematically review the literature of 35 IF-AT studies published from 1990 to the present. Based upon the meta-analysis, three research questions can be optimally answered and their implications of new findings are also addressed below. The first research question is about the effectiveness of IF-AT compared with traditional multiple-choice tests without corrective and immediate feedback. The second question is which IF-AT form is more effective than the other. The third question is about the potential moderator of IF-AT effectiveness.

The Overall Effectiveness of IF-AT

IF-AT offers an answer-until-correct answer form as a corrective and immediate feedback tool for students to enhance their learning in assessment. It also provides partial credit rubrics to assess students' partial knowledge. According to the meta-analysis, the overall effect of IF-AT was .581 in the fixed model and .618 in the random model, both in the range of a medium effect. However, the heterogeneity between the 35 studies was so severe that results were inconclusive. A moderator analysis was needed to explain heterogeneity. Aided by the FAT-PET-PEESE meta-regression analysis to correct the influence of publication bias effect, the effectiveness of IF-AT was .522 for the fixed model. This result is very close to the overall effect size of .53 reported by Bangert-Drowns et al. (1991).

It can be seen in Tables 7 and 8 that overall net effect sizes of the scratch-card IF-AT and the computer-based IF-AT were above a medium effect compared with the traditional multiple-choice tests without corrective and instant feedback. In terms of Cohen's criteria, the scratch-card IF-AT studies conducted in language arts and social sciences were more effective (with a large effect) than the computer-based IF-AT studies conducted in language arts and social sciences (with a medium effect). However, the net effect of the computer-based IF-AT studies conducted in mathematics and science is a medium effect; while the net effect of the scratch-card IF-AT studies conducted in mathematics and science is a small effect (see Table 9). Although effect sizes of IF-AT vary from small to large effects (.463~.981), it demonstrates that IF-AT is an effective measurement tool regardless of its testing forms. It is probably because IF-AT requires students to be more engaged in the meaningful answer-until-correct testing than traditional multiple-choice testing without corrective and immediate feedback for students.

Comparative Effectiveness between Scratch-card IF-AT and Computer-based IF-AT

Due to the salient heterogeneity between studies of the scratch-card IF-AT, the moderator analysis revealed that the IF-AT effectiveness between the scratch-card forms and the computer-based forms depended on the context of testing domain. Regardless of whether it is a fixed model or a random model, the scratch-card IF-AT was more effective than the computer-based IF-AT when the testing domain was language arts and social sciences, while the overall effect size of the computer-based IF-AT studies was larger in magnitude when the testing domain was mathematics and science, as supported by both the conventional meta-analysis and the FAT-PET-PEESE meta-regression analysis.

The foregoing effect interaction could explain the previous controversial findings between Epstein et al. (2002) and Smiley (2011). The former claimed that the scratch-card IF-AT was more effective. The latter argued that there was no significant difference between them. Both arguments are plausible since the dispute was caused by ignoring the interaction of effect sizes between IF-AT forms and testing domains. Why does the scratch-card IF-AT works better than the computer-based IF-AT in language arts and social sciences?

One possibility is that the scratch-card IF-AT can let students scratch off his/her answer as if scratching a lottery ticket. It is a scratch-and-win game, which sustains student's curiosity and interest. To simulate an instant lottery ticket, the answer keys in the computer-based IF-AT may be re-designed to appear similar to a lottery scratched box with an opaque coating.

Another possibility was addressed by Brosvic et al. (2006) and DiBattista et al. (2009). They explained that students seemed to display less anxiety and more motivation under the scratch-card IF-AT tests. This affective factor may make the scratch-card IF-AT more advantageous than the computer-based IF-AT in the context of language arts and social sciences. Regarding the lower effectiveness of IF-AT studies conducted in mathematics and science, there are two possible reasons. The task of assessment in mathematics and science seeks not only for memory retention, but also mainly for conceptual or relational understanding which requires less immediate feedback (Kulhavy et al., 1976); while a student's memory retention requires more immediate feedback (Kulik & Kulik, 1988; Hattie & Timperley, 2007). Moreover, it is not easy to conduct the computer-based IF-AT in the MS context because the process of calculation in mathematics and science is preferred by hand writing rather than by keyboard input (Bennett et al., 2008).

Besides, some previous research (Clariana, 1990; Shute, 2007) indicated that answer-until–correct (IF-AT) tended to make students more frustrated if the test was a difficult subject such as mathematics. Therefore, it would be inevitable to reduce the effectiveness of feedback (Clariana, et al, 2000). Consequently, the pooled effect size of IF-AT studies conducted in mathematics and science is less effective than the pooled effect size of IF-AT studies in language arts and social sciences.

Moderators and Conclusion

Generally speaking, the overall effect sizes of IF-AT in language arts and social sciences ranges from a large effect (>.8) to a medium effect(>.5), while in mathematics and science, they range from a small effect(<.5) to a medium effect(>.5), and the overall effect size varies slightly across IF-AT forms. Three moderating variables, testing domain, followed by testing form and participants, could best explain the major heterogeneity of IF-AT effectiveness.

Both the conventional meta-analysis and the FAT-PET-PEESE meta-regression analysis consistently indicate that IF-AT would work more effectively in the context of language arts and social sciences than in the context of mathematics and science. The effect size of IF-AT studies in language arts and social sciences was in the range of a large effect size, while in mathematics and science, it was in the range of a medium effect size (in terms of Cohen's criteria). The outcome is consistent with the previous findings (Kulik & Kulik, 1988; Hattie &Timperley, 2007).

The past evidence suggested that corrective and instant feedback could enhance students' learning through verification and elaboration (Shute, 2007). IF-AT can provide immediate corrective feedback as a verification means to reinforce students' memory. This may explain why IF-AT has worked quite effectively in the context of language arts and social sciences. Nevertheless, IF-AT has worked a little bit less effectively in the context of mathematics and science because more conceptual or relational understanding than memory retention is often required. We believe that elaboration feedback developed by Shute (2007) may be of great help to improve the effectiveness of IF-AT in the context of mathematics and science. Apparently, a computer-based IF-AT test is more flexible and easier to implement elaboration feedback in order to enhance students' conceptual understanding. Technology can be effective as a tutor, a teaching tool and a learning tool (Ross, Morrison, & Lowther, 2010) so that it can make learning or testing more engaging, more meaningful and more interesting. Viewed in this light, IF-AT can let students becoming more actively involved in the testing process through a corrective, elaborated, and instant feedback on an item by item basis. Therefore, aligning assessment via computer-based IF-AT with learning is feasible and effective.

REFERENCES

Anderson, R. C., Kulhavy, R. W., & Andre, T. (1971). Feedback procedures in programmed instruction. *Journal of Educational Psychology, 62*(2), 148.

Bangert-Drowns, R. L., Kulik, C. L. C., Kulik, J. A., & Morgan, M. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research, 61*(2), 213-238.

Beevers, C. E., Wild, D. G., McGuine, G. R., Fiddes, D. J., & Youngson, M. A. (1999). Issues of partial credit in mathematical assessment by computer. *Association for Learning Technology Journal, 7*(1), 26-32.

Bennett, S., Maton, K., & Kervin, L. (2008). The ‘digital natives’ debate: A critical review of the evidence. *British Journal of Educational Technology, 39*(5), 775-786.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, U.K: John Wiley & Sons..

Borenstein, M. (2009). Effect sizes for continuous data. In H. Cooper, L. V. Hedges and J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 279-293). New York: Russel Sage Foundation.

Bowman, T. G., & Laurent, T. (2011). Immediate feedback and learning in athletic Training education. *Athletic Training Education Journal, 6*(4), 202-207.

Brosvic, G. M., Epstein, M. L., Dihoff, R. E., & Cook, M. J. (2006). Acquisition and retention of Esperanto: The case for error correction and immediate feedback. *The Psychological Record, 56*(2), 205.

Butler, A. C., Godbole, N., & Marsh, E. J. (2013). Explanation feedback is better than correct answer feedback for promoting transfer of learning. *Journal of Educational Psychology, 105*(2), 290.

Butler, A. C., Karpicke, J. D., & Roediger III, H. L. (2007). The effect of type and timing of feedback on learning from multiple-choice tests. *Journal of Experimental Psychology: Applied, 13*(4), 273.

Carels, E. J. (1975). The effects of false feedback, sex, and personality on learning, retention, and the Zeigarnik effect in programmed instruction. *Dissertation Abstract International, 36,* 2094A. (University Microfilms No. 75-22, 345).

Carmichael, J. (2009). Team-based learning enhances performance in introductory biology. *Journal of College Science Teaching, 38*(4), 54.

Clariana, R. B. (1990). A comparison of answer until correct feedback and knowledge of correct response feedback under two conditions of contextualization. *Journal of Computer-Based Instruction*, *17*(4), 125-129.

Clariana, R. B., Ross, S. M., & Morrison, G. R. (1991). The effects of different feedback strategies using computer-administered multiple-choice questions as instruction. *Educational Technology Research and Development, 39*(2), 5-17.

Clariana, R. B., Wagner, D., & Roher Murphy, L. C. (2000). Applying a connectionist description of feedback timing. *Educational Technology Research and Development, 48*(3), 5-22.

Cohen, J. (1992). Statistical power analysis. *Current Directions in Psychological Science, 1*(3), 98-101.

Corbett, A. T., & Anderson, J. R. (2001). Locus of feedback control in computer-based tutoring: Impact on learning rate, achievement and attitudes. In *Proceedings of ACMCHI 2001 Conference on Human Factors in Computing Systems* (pp. 245-252). New York: Association for Computing Machinery Press

DiBattista, D. (2005). The immediate feedback assessment technique: A learner-centered multiple-choice response form. *The Canadian Journal of Higher Education, 35*(4), 111.

DiBattista, D., & Gosse, L. (2006). Test anxiety and the immediate feedback assessment technique. *The Journal of Experimental Education, 74*(4), 311-328.

DiBattista, D., Gosse, L., Sinnige-Egger, J. A., Candale, B., & Sargeson, K. (2009). Grading scheme, test difficulty, and the immediate feedback assessment technique. *The Journal of Experimental Education, 77*(4), 311-338.

Dihoff, R. E., Brosvic, G. M., Epstein, M. L., & Cook, M. J. (2004). Provision of feedback during preparation for academic testing: Learning is enhanced by immediate but not delayed feedback. *The Psychological Record, 54*(2), 207.

Doucouliagos, H., Stanley, T. D., & Viscusi, W. K. (2014). Publication selection and the income elasticity of the value of a statistical life. *Journal of Health Economics, 33*, 67-75.

Epstein Enterprises (2017, July 7). Immediate feedback assessment technique. *Retireved from: http://www.epsteineducation.com/home/demo/demo1.htm*

Epstein, M. L., Brosvic, G. M., Costner, K. L., Dihoff, R. E., & Lazarus, A. D. (2003). Effectiveness of feedback during the testing of preschool children, elementary school children, and adolescents with developmental delays. *The Psychological Record, 53*(2), 177.

Epstein, M. L., Epstein, B. B., & Brosvic, G. M. (2001). Immediate feedback during academic testing. *Psychological reports, 88*(3), 889-894.

Epstein, M. L., & Brosvic, G. M. (2002). Students prefer the immediate feedback assessment technique. *Psychological reports, 90*(3), 1136-1138.

Epstein, M. L., Lazarus, A. D., Calvano, T. B., Matthews, K. A., Hendel, R. A., Epstein, B. B., & Brosvic, G. M. (2002). Immediate feedback assessment technique promotes learning and corrects inaccurate first responses. *The Psychological Record, 52*(2), 187-201.

Farland, M. Z., Barlow, P. B., Levi Lancaster, T., & Franks, A. S. (2015). Comparison of answer-until-correct and full-credit assessments in a team-based learning course. *American Journal of Pharmaceutical Education, 7*9(2), 21.

Finn, B., & Metcalfe, J. (2010). Scaffolding feedback to maximize long-term error correction. *Memory and Cognition, 38*(7), 951-961.

Fitch, E. F., & Hulgin, K. M. (2008). Achieving inclusion through CLAD: collaborative learning assessment through dialogue. *International Journal of Inclusive Education, 12*(4), 423-439.

Ghani, U. (2014, June). Effect of feedback mechanisms on students' learning in the use of simulation-based training in a computer engineering program. Paper presented at *121st ASEE Annual Conference and Exposition*, Indianapolis, IN.

Hattie, J. (2008).*Visible learning: A Synthesis of Over 800 Meta-Analyses Relating to Achievement*. New York, NY: Routledge.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81-112.

Heald, H. M. (1970). The effects of immediate knowledge of results and correlation of errors and test anxiety upon test performance. *Dissertation Abstract International, 31,* 1621A. (University Microfilms No. 70-17, 724).

Hedges, L.V., and Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.

Higgins, J.,Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analysis. *British Medical Journal, 327*, 557-560.

Hunter, J. E. & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). Thousand Oaks: Sage Publications.

Jones, D. D. (2004). *The use of feedback in web-based instruction: achievement, feedback study time, and efficiency*. (Unpublished doctoral dissertation). University of North Carolina at Wilmington, North Carolina, USA.

Kehrer, P., Kelly, K. & Heffernan, N. (2013). Does Immediate Feedback While Doing Homework Improve Learning. In Boonthum-Denecke, Youngblood(Eds) *Proceedings of the Twenty-Sixth International Florida Artificial Intelligence Research Society Conference*, FLAIRS 2013, St. Pete Beach, Florida. May 22-24, 2013. AAAI Press 2013. p 542-545.

Kluger, A. N., & Denisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*(2), 254–284.

Kulhavy, R. W., Yekovich, F. R., & Dyer, J. W. (1976). Feedback and response confidence. *Journal of Educational Psychology, 68*, 522-528.

Kulik, J. A., & Kulik, C. L. C. (1988). Timing of feedback and verbal learning. *Review of Educational Research, 58*(1), 79-97.

Li, M., N. (2015). *Theory and practice of classical Meta-analysis: ESS and Excel*. Taipei, Taiwan: Wunan Publisher.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

Lopez, L. (2009). *Effects of delayed and immediate feedback in the computer-based testing environment.* ProQuest Dissertations Publishing, 2009. 3358462, Indiana State University.

Marsh, E. J., Lozito, J. P., Umanath, S., Bjork, E. L., & Bjork, R. A. (2012). Using verification feedback to correct errors made on a multiple-choice test. *Memory, 20*(6), 645-653.

Maurer, T. W., & Kropp, J. J. (2015). The impact of the Immediate Feedback Assessment Technique on course evaluations. Teaching and Learning Inquiry: *The ISSOTL Journal, 3*(1), 31-46.

Merrel, J. D., Cirillo, P. F., Schwartz, P.M. & Webb, J. A. (2015). Multiple choice testing using immediate feedback – assessment technique (IF AT®) forms: assessing learning from mistakes . *Higher Education Studies, 5*(5), 50-55.

Mohrweis, L. C., & Shinham, K. M. (2015). Enhancing students' learning: Instant feedback cards. *American Journal of Business Education (Online), 8*(1), 63.

Morrison, G. R., Ross, S. M., Gopalakrishnan, M., & Casey, J. (1995). The effects of feedback and incentives on achievement in computer-based instruction. *Contemporary Educational Psychology, 20*(1), 32-50.

Murphy, P. (2007). Reading comprehension exercises online: The effects of feedback, proficiency and interaction. *Language Learning and Technology, 11*(3), 107-129.

O'Neil, M., Rasor, R. A., & Bartz, W. R. (1976). Immediate retention of objective test answers as a function of feedback complexity. *Journal of Educational Research, 70,* 72–75.

Olson, G. H. (1971). A multivariate examination of the effects of behavioral objectives, knowledge of results and assignment of grades on facilitation of classroom learning. *Dissertation Abstract International, 32,* 6214A. (University Microfilms, No.72-13, 552).

Peck, S. D., Werner, J. L. S., & Raleigh, D. M. (2013). Improved class preparation and learning through immediate feedback in group testing for undergraduate nursing students. *Nursing Education Perspectives, 34*(6), 400-404.

Persky, A. M., & Pollack, G. M. (2008). Using answer-until-correct examinations to provide immediate feedback to students in a pharmacokinetics course. *American Journal of Pharmaceutical Education, 72*(4), 83.

Peters, S. U. (2015). *Exploring the effectiveness of collaborative assessment preparation with immediate feedback in an intensive adult English as a second language classroom*. (Unpublished doctoral dissertation), The Florida State University, Florida, USA.

Pressey, S. L. (1926). A simple apparatus which gives tests and scores-and teaches. *School and Society, 23*(586), 373-376.

Pridemore, D. R., & Klein, J. D. (1991). Control of feedback in computer-assisted instruction. *Educational Technology Research and Development, 39*(4), 27-32.

Prieto, G., Velasco, A. D., Arias-Barahona, R., Anido, M., Núnez, A. M., & Có, P. (2010). Training of Spatial Visualization Using Computer Exercises. *Journal for Geometry and Graphics, 14*(1), 105-115.

Razagifard, P., Ghabelnezam, A., & Fard, V. R. (2011) The effect of computer-mediated feedback on second language reading comprehension. *International Journal on New Trends in Education and Their Implications, 2*(1), 7.

Robinson, D. H., Sweet, M., & Mayrath, M. (2008). A computer-based, team-based testing system. In D. H. Robinson, J. M. Royer & G. Schraw (Series Eds.), *Recent innovations in educational technology that facilitate student learning current perspectives on cognition, learning and instruction* (pp. 277–290).

Rosa, E. M., & Leow, R. P. (2004). Computerized task-based exposure, explicitness, type of feedback, and Spanish L2 development. *The Modern Language Journal, 88*(2), 192-216.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin, 86*(3), 638.

Rosnow, R. L., Rosenthal, R., & Rubin, D. B. (2000). Contrasts and correlations in effect-size estimation. *Psychological Science, 11*(6), 446-453.

Ross, S. M., Morrison, G. R., & Lowther, D. L. (2010). Educational technology research past and present: Balancing rigor and relevance to impact school learning. *Contemporary Educational Technology, 1*(1), 17-35.

Samuels, S. J., & Wu, Y. (2003). *The effects of immediate feedback on reading achievement*. Unpublished manuscript, University of Minnesota, Minneapolis. Retrieved from http://www.tc.umn.edu/~samue001/web pdf/immediate_feedback.pdf.

Schneider, J. L., Hein, S. M., & Murphy, K. L. (2014). Immediate answer-until-correct feedback in chemistry testing. *Biennial Conference on Chemical Education*. Retrieved from: https://docs.google.com/viewer?url=http://www.enfusestem.org/wpcontent/ uploads/ 2 016/04/2016NSF-Symposium-Poster-2016_04_11.pdfandembedded=true.

Shute, V. J. (2007). Focus on Formative Feedback. *ETS Research Report Series, 2007*(1), i-47. doi:doi:10.1002/j.2333-8504.2007.tb02053.x

Skinner, B. F. (1954). The science of learning and the art of teaching. *Harvard Educational Review, 24*, 86-97.

Slepkov, A. D. (2013). Integrated testlets and the immediate feedback assessment technique. *American Journal of Physics, 81*(10), 782-791.

Slepkov, A. D., Vreugdenhil, A. J., & Shiell, R. C. (2016). Score increase and partial-credit validity when administering multiple-choice tests using an answer-until-correct format. *Journal of Chemical Education*, *93*(11), 1839-1846.

Smiley, W. F. (2011). *A systematic evaluation of the immediate feedback assessment technique*. (Unpublished master thesis). James Madison University, Virginia, USA.

Spiegelhalter,D. J., Abrams, K.R., & Myles, J.P. (2004). *Bayesian approaches to clinical trials and health-care evaluation*. Chichester, England: John Wiley & Son.

Stanley, T. D. (2008). Meta-regression methods for detecting and estimating empirical effects in the presence of publication selection. *Oxford Bulletin of Economics and statistics*, *70*(1), 103-127.

Stanley, T. D., & Doucouliagos, H. (2012). *Meta-regression analysis in economics and business*. London and New York: Routledge.

Stanley, T. D. (2017). Limitations of PET-PEESE and Other Meta-Analysis Methods. *Social Psychological and Personality Science*, 1948550617693062. Retrieved from: DOI: https://doi.org/10.1177/1948550617693062.

Sterne, J. A., Becker, B. J., & Egger, M. (2005). "*The funnel plot" in publication bias in meta-analysis: Prevention, assessment and adjustments* (*eds* H.R. Rothstein, A.J. Sutton andM. Borenstein), 75-98. Chichester: John Wiley and Sons.

Thompson, L. G. (1975). *A study of the effect of an answer-until-correct multiple-choice procedure on mathematics achievement*. (Unpublished doctoral dissertation). Oregon State University, Oregon, USA.

Valdez, A. (2012). Computer-based feedback and goal intervention: learning effects. *Educational Technology Research and Development*, *60*(5), 769-784.