6 月 6 日

Barkaoui, K. (2011). Think-aloud protocols in research on essay rating: An empirical study of their veridicality and reactively. *Language Testing, 28,* 51-75.

1. Introduction

■Think-aloud protocols (TAPs) have been widely used to investigate, and build models of, essay rating processes in L1 and L2 context.

■Advantages of TAPs

・In the context of essay rating, TAPs can help uncover participants' thought processes to explain why and how a rater arrives at a given score.

・Compared to other self-report method (interviews and questionnaires), TAPs have the added advantage of being immediate, thus avoiding problems of information retrieval and/or filtering.

・TAPs are more likely to reflect what raters actually do and are concerned about as they read and rate essays, rather than what they believe they do and are concerned about as is usually revealed in other self-report methods.

・While other self-report methods provide generalized statements about behavior, TAPs inspect specific instances of actual behavior.

■Limitations of TAPs

・TAPs have several limitations, but the main criticism of it concerns their veridicality and reactivity.

Veridicality: Whether the TAPs accurately report and represent the participants' true and complete thinking and rating process.

Reactivity: Whether the requirement to report the rating process alters the process being observed and/or its outcomes.

■In this thesis, 16 previous think-aloud studies are summarized (see Table 1 pp. 53-56).

➜ While several researches have used TAPs to investigate essay rating processes, very few empirically examined the veridicality and reactivity of this technique.

➜None of the studies collected or reported empirical data about whether and how TAPs affected their participants' rating processes.

➜None of the studies examined whether TAP effects, and the quality and quantity of participant' verbalizations, across individual and context.

■The goal of this study was to assess the veridicality and reactivity of TAPs across rater groups (novice vs. experienced) and rating scales (holistic vs. analytic).

■Specifically this study addressed the following questions:

RQ1: How do raters perceive the completeness and accuracy of their TAPs? Do raters' perceptions of the completeness of their TAPs vary across rater groups and rating sales?

RQ2: Does the requirement to think aloud affect the processes and outcomes (i.e., scores) of essay

rating? If yes, how and to what extent? Do these effects vary across rater groups and rating scales?

2. Method

2.1 Participants: 6o raters (31 novice and 29 experienced)

Of the 6o raters, 25 raters (11 novice and 14 experienced) provided TAPs and remaining 34 participants (19 novice and 15 experienced) rated all the essays silently.

Table 1 Typical profile of a novice an experienced rater

	Novice	Experienced
Role at time of the research	Student in TESL program	ESL Teacher
ESL teaching experience	None	10 years or more
Experience teaching ESL Writing	None	5 years or more
ESL essay rating experience	None	5 years or more
Received training in assessment	No	Yes
Post-graduate study	None	M.A./M.Ed.
Professional certificate	TESL in progress	TESL Certificate

2.2 Procedures

■There were two phases in this experiment, and participants rated essays using two scales (holistic/analytic) in different conditions (silent or TA).

■The holistic and analytic scales differed in terms of whether to assign one overall score (holistic) or multiple scores (analytic) to each essay. The analytic scale included five rating dimensions: communicative quality, organization, argumentation, linguistic accuracy, and linguistic appropriacy.

■Phase 1:

1. Participant attended orientation session about scale 1
2. Each silent rater rated 24 essays, while each think-aloud rater rated 12 essays, silently using scale 1 (at home).
3. Think-aloud raters received training on thinking aloud.
4. Each think-aloud rater rated 12 essays while thinking aloud using scale 1 (at home).
5. Participants responded to interview about rating scale 1 and process.

■Phase 2: (after two weeks)

6. Participant attended orientation session about scale 2.
7. Each silent rater rated 24 essays, while each think-aloud rater rated 12 essays, silently using scale 2 (at home; the same essays as in 2 above).
8. Each think-aloud rater rated 12 essays while thinking aloud using scale 2 (at home; the same essays as in 4 above).
9. Participant responded to interview about rating scale 2 and process and TAPs.

2.3 Data sources and analysis

■Three strategies, that combined score analyses and self-report data, were adopted in this study.

①A multifaceted Rasch measurement model (MFRM), using FACET (Linacre, 2007), was used to compare rater severity and self-consistency under TA (Think-Aloud) and non-TA conditions across and within raters,

②The second strategy consisted of examining the think-aloud protocols themselves for evidence and explanations of TAPs veridicality and reactivity

➜Two types of TAP-related comments were identified:

(a) Explicit comments: comment explicitly related to thinking aloud

(b) Implicit comments: comments that were identified impressionistically and related mainly to the social and interactive aspects of TAPs.

③The third strategy was to interview the participants about their perceptions of thinking-aloud and its effects.

➜The TA related comments were classified in terms of the following general scheme:

1. Veridicality and Completeness

(a) Omission, (b) Commission, (c) participant' explanation for (a) and (b)

2. Reactivity

Comments concerning whether the act of thinking aloud affected the participants' rating processes and/or outcomes fall under this category.

■The results were then compared across rating scales, raters and rater groups.

3. Findings

■Raters' perceptions of veridicality of their TAPs

・More than a third of the participants (n = 9) reported that they were not able to verbalized all their thoughts during the TAPs.

・The majority of the participants (n = 6) who felt that they could not possible report all their thoughts were experienced raters.

・The responses of some participants (n = 4) suggest that the rating scales affected what and how much they reported.

■TAPs reactivity

□TAP effects on essay scores

・To examine the effects of TA on rater severity, a FACETS bias analysis of rater-by-rating condition interactions was conducted.

・The Pearson r correlation between FACETS estimates of rater severity across the two rating condition (silent/TA) was .83. However, FACETS detected 10 significantly biased rater-bu-rating condition interactions out of 50 possible interactions.

→There was an interaction effect between rating condition and rating scale, but this effects was not statically significant.

・Effects of TA on rater self-consistency were examined by means of rater fit statistics.

→There was a slight increase in terms of the average infit mean square statistics from .92 *(SD* = .40) under the silent condition to 1.06 (*SD* = .42) under the TA condition.

□Raters' perceptions of TAP effects on their rating processes

・Fourteen participants reported that rating while TA seems to have taken longer.

・Seven participants felt that TA may have changed how they interacted with the essays and how they performed the rating tasks.

・The rater's overall approach to essay rating were affected by TA.

・Seven participants reported that TA lowered their confidence in their rating.

・Some participant felt that TA led them to consider the essays and the ratings more carefully (*n* = 4) and/or to produce more explanations and justifications of their ratings (*n* = 5).

・Six participants reported several types of effects of TAPs on their rating criteria, but majority of participants (n = 19) reported that TAPs did not affect the scores they assigned.

□Raters' explanation of TAP effects

・Most of the participants (*n* = 16) felt that thinking aloud was "very difficult", "tiring" and "very demanding".

・Six participants reported that hearing their own voice while rating made them second guess themselves, while others reported that it made them process and judge the essays better.

・TA drew and increased some participants' (*n* = 9) attention to their rating processes which some found positive and others negative.

・Although the participants were specifically instructed to "talk and act as if you are talking to yourself" during the TAPs, awareness of an audience for the protocols seems to have affected the performance and verbalization of several participants (*n* = 7).

4. Discussions

■The findings of this study indicate that TAPs were incomplete and that they altered the rating processes, severity and self consistency of some participants.

■TAPs seem to affect various aspects of the rating process, rating criteria and aspects of writing attended to, decision making processes rater confidence, as well as rater severity and self-consistency.

■The findings also indicate that TAP incompleteness and effects depend on other individual and contextual factors as well such as rater characteristics and rating scale type.

■Reading aloud and reading silently focus raters' attention on different aspects of writing and, as a result, alter the rating process and outcomes. Reading aloud seems t focus attention and enhance ability to detect micro-level problems in the essays, but hinders global essay comprehension.

■The result highlight the important role that social factors, specifically audience awareness, play in TAPs. It is crucial that researchers take into consideration the fact that TAPs are socially and interactively constituted when collecting, analyzing and interpreting TAP data.

5. Limitations

・TA and silent ratings were not counterbalanced.

・The essays rated under the two conditions were different.

・None of the participants had experience with the rating scales used in this study.

・That English was not the first language of some participants might have affected the quantity and quality of their verbalizations.

・Although both scores and protocols were closely examined to address the research questions, the main data for this study came from interviews with raters about their perceptions of the completeness and effects of TAPs.

➜Finding from this study need to be interpreted with caution and treated as hypotheses for further research.

6. Implications

・TAPs are necessarily incomplete and likely to alter the rating process.

・The limitations do not mean that TAPs should not be used in research on essay rating processes.

- TAPs are probably the only tool to provide some insight, though incomplete and imprecise, into the kinds of processes that raters employ to complete rating tasks, including their evaluation criteria and decision-making behaviors, the conflicts they face and the strategies they employ to resolve them.

- Reactivity and incompleteness are not unique to TAPs.

-The finding that protocols are dialogic and socially situated is no unique to this technique; other approaches, such as interviews and experiments, also have several social and interactive features.

・

・